FORMATIVE LANGUAGE ASSESSMENT FOR ENGLISH LEARNERS – AN EXPLORATION OF VOCABULARY DIVERSITY INDICES IN WRITING CBM

by

BRITT CHRISTENSEN LANDIS

A DISSERTATION

Presented to the Department of Special Education and Clinical Sciences
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

June 2021

DISSERTATION APPROVAL PAGE

Student: Britt Christensen Landis

Title: Formative Language Assessment for English Learners – An Exploration of Vocabulary Diversity Indices in Writing CBM

This dissertation has been accepted and approved in partial fulfillment of the requirements
for the Doctor of Philosophy degree in the Department of Special Education and Clinical Sciences by:

Dr. Ben Clarke	Co-Chair
Dr. Sylvia Linan-Thompson	Co-Chair
Dr. Patrick Kennedy	Member
Dr. Audrey Lucero	Member

and

Kate Mondloch	Interim Vice Provost and Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded June 2021.

DISSERTATION ABSTRACT

Britt Christensen Landis

Doctor of Philosophy

Department of Special Education and Clinical Sciences

June 2021

Title: Formative Language Assessment for English Learners – An Exploration of Vocabulary Diversity Indices in Writing CBM

None of the most well-established indices within curriculum-based measurement of writing (CBM-W) assess students' use of vocabulary within writing. Yet vocabulary knowledge is fundamental to writing, as well as the other three domains of English language proficiency (reading, speaking, listening), and it is particularly crucial for English learners (ELs). This dissertation study explored reliability, validity, and classification accuracy for two types of CBM-W vocabulary indices for ELs: Number of Different Words (NDW; a production-dependent index) and Corrected Type Token Ratio (CTTR; a production-independent index). One-hundred and seventy-five second grade ELs were administered six monthly CBM-W probes and a state-issued English Language Proficiency test, which included scores for all four language domains. Results were similar for both indices, but preliminary findings indicated that NDW was more reliable and had slightly stronger classification accuracy than CTTR. For NDW, delayed alternate form reliability was moderate. NDW criterion validity correlations were variable, but mostly moderate with the writing and reading criterion scores, and mostly weak with the listening and speaking scores. Classification accuracy was weak for overall English language proficiency status, but moderate with the writing domain, and to a lesser extent,

the reading domain. Results are contextualized within an MTSS framework for supporting the diverse language needs of ELs in second-grade.

CURRICULUM VITAE

NAME OF AUTHOR: Britt Christensen Landis

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene
Colorado College, Colorado Springs

DEGREES AWARDED:

Master of Science, Psychology, 2019, University of Oregon
Bachelor of Arts, Sociology, 2011, Colorado College

AREAS OF SPECIAL INTEREST:

Bilingual Education and Educational Practices for English Learners
Formative Assessment for Early Literacy and Mathematics
Explicit and Systematic Instruction for Early Literacy and Mathematics
Positive Behavior Supports

PROFESSIONAL EXPERIENCE:

School Psychologist Intern, Springfield Public Schools, Springfield, OR, 2020-2021
Graduate Employee, Center on Teaching and Learning (CTL), University of Oregon, Eugene, OR 2017-2020
Advanced Practicum Student, Center on Teaching and Learning (CTL), University of Oregon, Eugene, OR 2018-2020
Integrated Practicum Student, Bethel School District, Eugene, OR, 2017-2018
English as a Foreign Language Teacher, Universidad Sergio Arboleda, Bogotá, Colombia

PUBLICATIONS

Clarke, B., Sutherland, M., Doabler, C. T., Kelsey, N., & Landis, B. (2019). *Developing and investigating the promise of early measurement screeners*. Manuscript submitted for publication to School Psychology Review.
Scalise, K., Irvin, P. S., Alresheed, F., Zvoch, K., Yim, H., Park, S., Landis, B., Meng, P., Kleinfelder, B., Halladay, L., & Partsafas, A. (2017). *Accommodations in digital interactive STEM assessment tasks: Current accommodations and promising practices for enhancing accessibility for students with disabilities*. Manuscript accepted for publication in the *Journal of Special Education Technology*.

ACKNOWLEDGEMENTS

I am deeply thankful for the guidance of my dissertation chairs, Dr. Ben Clarke and Dr. Sylvia Linan Thompson, and to committee members Dr. Patrick Kennedy and Dr. Audrey Lucero. I am incredibly grateful to have had Dr. Clarke's thoughtful, supportive, and down-to-earth mentorship throughout my studies at the University of Oregon. I also thank Dr. Linan Thompson and Dr. Kennedy for including me in Project Write. Both have been enormously generous: Dr. Linan Thompson with her incredible knowledge of bilingual literacy supports, and Dr. Kennedy with his expertise and patience as he supported me with data analysis. I also thank Marissa Pilger Suhr, for being my primary thought partner throughout this program and a wonderful friend. Lastly, I'm forever grateful to my partner Daniel, and all my family and friends, for their many years of love, cheering, adventures, and kindness. The research reported here was supported in part by the National Institutes of Health, U.S. Department of Health and Human Services, awarded to the University of Oregon through grant 5R21HD095487-02. The principal investigators for this grant were Dr. Sylvia Linan-Thompson and Dr. Patrick Kennedy.

TABLE OF CONTENTS

Chapter Page

Chapter Page

LIST OF FIGURES

Figure Page

LIST OF TABLES

Table Page

I. INTRODUCTION

At school, language is heard in the hallway and spoken by a student raising her hand. It is read in a poem and in a chapter on human migration. Language is written to tell a story and to explain a math procedure. Language is both central to life at school and one of school's most important goals. Ensuring that students have strong language skills across all four modalities - listening, speaking, reading, and writing – is fundamental to preparing students for engaged citizenship, higher education, and work (Baker, Simmons, and Kame'enui, 1995; Council of Chief State School Officers, 2014; DiCerbo, Anstrom, Baker, & Rivera, 2014; Koutsoftas, 2013; Nuñez, Rios-Aguilar, Kanno, & Flores, 2016; Troia, 2009). For English learners (ELs), access to strong language skills in English is especially important. ELs are emergent bilingual students with home languages other than English and who have not yet developed English Language Proficiency (ELP; de Brey et al., 2019; Goldenberg & Coleman, 2010; Goldenberg, Reese, & Rezaei, 2011).

Access to English Language Proficiency

Schools are legally obligated to ensure that ELs have the opportunity to develop ELP - which is defined as having the listening, speaking, reading, and writing skills (the four forms or domains of language) necessary to fully benefit from and participate in grade level coursework in English at school, without scaffolding and support (Gottlieb, 2016; U.S. Department of Education Office for Civil Rights, 2015). Emergent bilinguals who are allowed this opportunity and become proficient in English (former ELs) don't experience the academic achievement gaps that are consistently found between non-ELs and current ELs (de la Torre, Blanchard, Allensworth, & Freire, 2019). Former ELs score significantly better than current ELs and *as well or better than* native English speakers

across a range of important educational outcomes – including high stakes achievement tests and high school graduation rates (de la Torre et al., 2019; New York State Education Department, 2016; Oregon Department of Education, 2019).

English learners have widely varying histories of language exposure and language abilities and thus need varying levels of support to develop strong English language skills (Goldenberg & Coleman, 2010; Hammer et al., 2014; Iglesias & Rojas, 2012). Though developing ELP takes time for all learners, evidence suggests that too many students are never exited from EL status and are not given the opportunity to develop strong English language skills (de la Torre et al., 2019; Hakuta, Butler, & Witt, 2000; Umansky, 2016). For instance, in one of the few longitudinal studies monitoring differing EL trajectories, de la Torre et al. (2019) found that about 25% of Chicagoan ELs didn't develop ELP by fifth grade and were unlikely to reach ELP later – only 5% of them exited EL status by 8th grade. Additionally, de la Torre et al. (2019) found that the students who didn't develop ELP by 8th grade had started with lower English Language Proficiency scores in K-3rd grade. This suggests that students who are on not on track for developing ELP can be identified early and thus, provided with intervention. Much more research is needed to both identify students who need more English language support and evaluate the practices supporting them.

MTSS and Formative Assessment of Language

To ensure that all ELs have the opportunity to develop ELP, multi-tiered systems of support (MTSS) are necessary. MTTS is increasingly recognized as best practice, especially in elementary schools, for meeting the diverse needs of all students (Gersten et al., 2009; Pullen et al., 2010). MTSS is designed to provide varying levels of support,

depending on student need. More specifically, this involves providing high quality instruction to all students, identifying students who are at-risk, and providing interventions to at-risk students (Gersten et al., 2009; Kettler, Glover, Albers, & Feeney-Kettler, 2014; VanDerHeyden & Burns, 2018). One essential element of this model is formative assessment data. Formative assessment data can be used to screen large groups of students efficiently and determine which students need more support, evaluate instructional programming based on screening data, provide diagnostic information about which skills need targeting for individual students, and monitor growth and response to evidence-based interventions (Cummings, Stoolmiller, Baker, Fien, & Kame'enui; Deno, 2003; Gersten et al., 2009; Kettler et al., 2014). However, although formative assessments of reading are widely researched and published, assessment of other areas of language, such as writing, are less well developed for elementary school aged students. For ELs, this is particularly problematic because their EL status is contingent upon their performance across all four language domains. It is rare that formative assessment researchers address the specific needs and characteristics of emergent bilingual ELs, and adequately represent them in their research (Campbell, Espin, & McMaster, 2013; Keller-Margulis, Payan, Jaspers, & Brewton, 2016).

Gaps in Formative Language Assessment

Another barrier for ELs is that formative assessments of language for students in elementary school tend to focus on some component skills of language more than others. Most formative assessment research focuses on the skills most related to deciphering and using the alphabetic code of written language, such as decoding, fluent word recognition, spelling, and handwriting. On the other hand, fewer formative language assessment

studies focus on the oral language skills most related to comprehending and conveying ideas through language, such as vocabulary, listening comprehension, and oral proficiency (Adlof & Hogan, 2019; Biemiller, 2012; Catts, Hogan, & Adlof, 2005; National Center on Intensive Intervention, n.d.; Pearson, Hiebert, & Kamil, 2012; Scott, 2009; Smith & Lembke, 2020). This lack of research on how to formatively assess oral language skills is problematic for the sizable group of students whose reading and writing difficulties are due, at least in part, to inadequate oral language skills (Adlof & Hogan, 2019; Babayiğit, 2014; Catts et al., 2005; Juel, Griffith, & Gough, 1986; Lesaux, Crosson, Kieffer, & Pierce, 2010; Lesaux & Kieffer, 2010; Mancilla-Martinez & Lesaux, 2011). For ELs, the lack of research on oral language skills is especially problematic because they are more likely than their non-EL peers to need additional support with these skills in English (August, Carlo, Dressler, & Snow, 2005; Lesaux & Kieffer, 2010; Lesaux et al., 2010; Mancilla-Martinez & Lesaux, 2011). The present study will place a particular emphasis on addressing the research gap on measures of oral language skills at the *word* level, or in other words – vocabulary. Vocabulary takes center stage because it is widely recognized as one of the most fundamental components of dual and second language acquisition (August et al., 2005; Mancilla-Martinez & Lesaux, 2011 2005; Manyak, 2012; Read, 2000).

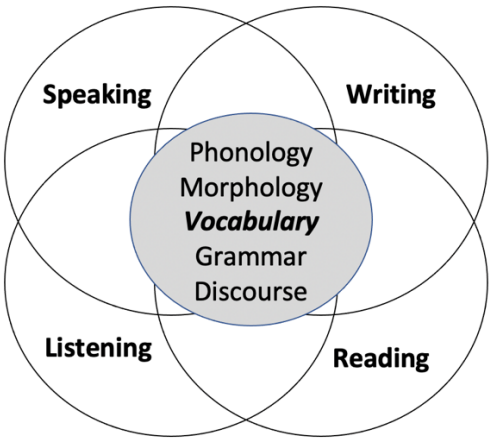
Importance of Vocabulary to Language Development

Vocabulary is implicated in all four of the listening, speaking, reading, and writing domains of language (Koutsoftas, 2013). A conceptual model of language depicted in Figure 1 (adapted from Koutsoftas, 2013) shows the four overlapping domains of language, with both unique and shared elements. For example, though code-

related skills are important to both the reading and writing domains, vocabulary is fundamental to all four language domains. Therefore, vocabulary is depicted at the center of the conceptual diagram in Figure 1, along with phonology, morphology, grammar, and discourse (Gottlieb, 2016; Koutsoftas, 2013; Moats, 2000; Shanahan, 2006; WIDA Consortium, 2012).

Figure 1

Conceptual Diagram of Language, Adapted from Koutsoftas, 2013



Correlational research, including longitudinal studies, support vocabulary's positioning in this conceptual model. Vocabulary is consistently found to be correlated with and predictive of both the oral language domains (listening and speaking) and written language domains (reading and writing) for monolingual and bilingual learners (Hwang, Mancilla-Martinez, Mcclain, Oh, & Flores, 2019; Gottlieb, 2016; Koutsoftas, 2013; Kieffer, 2012; Kim, 2016; Kim & Schatschneider, 2017; Lesaux et al., 2010; Lonigan & Milburn, 2017; Moats, 2000; Silverman et al., 2015; Simon-Cereijido & Gutierrez-Clellen, 2009; Staehr, 2008; WIDA Consortium, 2012; Zareva, Schwanenflugel, & Nikolava, 2005). Vocabulary and broader language outcomes are also causally related. For instance, multiple nationally funded comprehensive literature

reviews found strong evidence that vocabulary instruction improves reading comprehension (the domain in which most experimental vocabulary research has been conducted), including the National Reading Panel (NRP) review in 2000 and an Institute of Education Science's What Works Clearinghouse Guide (Foorman et al., 2016; NRP, 2000). There is also considerable evidence that teaching vocabulary improves language outcomes for ELs specifically. For instance, in a comprehensive literature review, Baker et al. (2014) found strong evidence that vocabulary-focused instruction improves literacy and academic content knowledge for ELs.

Particular Importance of Vocabulary for English Learners

Though vocabulary knowledge is crucial to academic success for all learners, it is particularly important for English learners for several reasons. One reason for this is that vocabulary is more language specific than other language-related skills. ELs have lower English vocabulary knowledge than non-ELs but this is not often the case for many other language-related skills (August et al., 2005; Dressler & Kamil, 2006; Lesaux et al., 2010; Mancilla-Martinez & Lesaux, 2011; Manis, Lindsey, & Bailey, 2004). For instance, phonological awareness and word reading ability appear to transfer between languages relatively easily (at least within languages with similar alphabets like English and Spanish), and ELs tend not to need more support than non-ELs in these areas (August et al., 2005; Dressler & Kamil, 2006; Mancilla-Martinez & Lesaux, 2011; Manis, Lindsey, & Bailey, 2004). Similarly, ELs also tend to apply discourse related skills such as story structure, cohesion, and comprehension monitoring strategies across languages and tend to score similarly to non-ELs in these areas as well (Dressler & Kamil, 2006; Manis et al., 2004). Though there are aspects of vocabulary knowledge that can also be leveraged

across languages, such as underlying content knowledge and cognates, emergent bilinguals still face a significant challenge in developing sufficient vocabulary knowledge in both languages and often need targeted support to help them meet this challenge (August et al., 2005; Dressler & Kamil, 2006; Lugo-Neris, Jackson, & Goldstein, 2010; Mancilla-Martinez & Lesaux, 2011).

A second, related reason that attention to vocabulary is especially critical for ELs is due to the "immense volume of information involved (Nagy & Herman, 1987, p. 20)." For instance, Nation (2006) estimated that readers need to have a vocabulary size of at least 8,000 to 9,000 word families to understand newspapers and novels. Additionally, learning vocabulary involves more than lists of word meanings; depth, or quality, of vocabulary knowledge is also critical (Beck, McKeown, & Kucan; Coyne, McCoach, Loftus, Zipoli, & Kapp, 2009; Nagy, Townsend, Lesaux, & Schmitt, 2012; Pearson, Hiebert, & Kamil, 2012; Read, 2000; Zareva et al., 2005). Even for English-only monolingual students, developing adequate word knowledge requires time and support – for English learners, attention to vocabulary is critical (August et al., 2005).

Promising Formative Assessment Approach: Vocabulary Use in CBM-Writing

As discussed above, research on formative measures of vocabulary and other oral language skills is less developed compared to measures of code-related skills. However, curriculum-based measurement of writing (CBM-W) may be one promising approach for addressing this gap. CBM-W is a well-established formative assessment framework for assessing written language. With CBM-W, students' authentic language samples are scored for a variety of language features with objective, analytic index scores. The most common CBM-W indices with the most validity evidence are words written (WW),

correctly spelled words (CSW), and correct word sequences (CWS), as well as variations of these indices (Benson & Campbell, 2009; Romig et al., 2017). However, none of the most common and well-established CBM-W indices focus on vocabulary (Smith & Lembke, 2020). This study explores two promising vocabulary indices that can be derived from CBM-W assessments: Number of Different Words (NDW) and Corrected Type Token Ratio (CTTR).

Before delving deeper into NDW and CTTR, and why these specific indices were chosen, it is first important to explore some implications for using a CBM-writing format. First, this assessment approach is intended to be practical for frequent and efficient use within schools –a core feature of formative assessment in general, and CBM in particular (Deno, 1985). In addition to being standardized, CBM-W tasks are brief, can be group administered, and can be hand scored without speech to text language analysis software. With CBM-W tasks, elementary school aged students are typically given about 3-5 minutes to write in response to a standardized picture, story starter, sentence starter, or question (Romig, Therrien, & Lloyd, 2017).

Implications of a Written Discourse Context

This study's assessment approach also differs substantially from a common traditional vocabulary assessment approach, in which students are tested on their ability to recognize or produce specific words that were selected by the test developer (Read, 2000; Read & Chapelle, 2001). With a CBM-W approach, students are tested on their ability to make use of the vocabulary knowledge they have to communicate effectively and fluently. One implication of this approach is that it is considered more authentic and

educationally relevant, and less culturally biased (American Speech-Language-Hearing Association, 2004; Deno, 1985; Wood, Wooford, Gabas, & Petscher, 2018).

Using written discourse as an assessment format also necessarily requires students to integrate their English vocabulary knowledge with other language skills. Writing is a complicated process that involves many skills (Benson & Campbell, 2009; Grobe, 1981; Kim & Schatschneider, 2017; McMaster & Espin, 2007; Juel et al., 1986). For instance, students need to be able to transcribe their ideas and vocabulary knowledge into written words, which also requires handwriting skills and proficiency with the alphabetic code (Juel et al., 1986; Kim & Schatschneider, 2017). The written discourse task furthermore requires students to integrate their vocabulary knowledge with other oral language skills to communicate their ideas, such as their knowledge of grammar, and higher-order discourse-related skills such as theory of mind and knowledge of text structures (Kim & Schatschneider, 2017; Koutsoftas, 2013; Read, 2000; Scott, 2009).

Implications of an Expressive Task

Another implication of a CBM-writing format is that it requires students to use expressive, or productive, language, which theoretically invokes a more advanced form of vocabulary knowledge than receptive tasks (Nagy et al., 2012; McKeown et al., 2012; Pearson et al., 2012; Perfetti & Adlof, 2012; Read, 2000; Zareva et al., 2005). It's possible that the more advanced nature of an expressive vocabulary task, such as writing, could be seen as a limitation, because it could lead to floor effects and lack of sensitivity, an insensitivity to growth, or the inability to serve as in an indicator of receptive skills (Hwang et al., 2019; Lee & Muncie, 2006; Lugo-Neris et al., 2010). Conversely, there is also reason to suspect that vocabulary measured within writing is indeed also indicative

of vocabulary knowledge applied to receptive domains of language. In previous research, expressive vocabulary measures have outperformed receptive measures in predicting not just comprehensive language outcomes for ELs (including all four language domains), but also receptive outcomes in particular, such as listening and reading comprehension (Hwang et al., 2019; Kieffer, 2012). For instance, Hwang et al. (2019) found that expressive vocabulary, and not receptive vocabulary, was a significant predictor of a comprehensive English language proficiency assessment and a reading comprehension assessment for second- and fourth-grade Els. They suggested that perhaps expressive vocabulary assessments are better indicators of the quality dimension of vocabulary knowledge (Hwang et al., 2019). However, it is important to note that the majority of this research on expressive measures has used speaking and not writing tasks, including the study from Hwang and colleagues (2019). More research is needed to understand the implications of measuring vocabulary via productive writing tasks (Read, 2000).

Types of Vocabulary Indices in Writing

Several variations of written vocabulary indices have been studied within writing assessment and CBM-W. For instance, early CBM-W studies from the Institute of Research on Learning Disabilities (IRLD) at the University of Minnesota included two vocabulary indices in their studies, which examined the extent to which various writing indices could serve as general outcome measures (GOMs), or proxies for, broader writing outcomes, for 3rd – 6th graders (McMaster & Espin, 2007). These studies were not targeted for ELs, but it is not known whether ELs were included in their samples. They selected indices that counted the number of mature, or advanced words that students wrote: Number of Long Words and Number of Mature words (McMaster & Espin, 2007) and

found variable, but generally moderate to strong validity coefficients for these indices (.29-.88). However, subsequent studies have shown less promising results for these and similar measures (Gansle et al., 2002; Olinghouse & Leaird, 2009). For instance, Gansle et al. (2002) found that Number of Long words and the WordPerfect formula (which also intends to measure the use of mature words) were among the least reliable measures that they tested.

Vocabulary Diversity Indices

Another promising approach to analyzing students' use of vocabulary in writing is to measure the extent to which students use a *range* of vocabulary in their writing. Indices that align with this approach are typically called measures of vocabulary diversity, or lexical diversity. The NDW and CTTR indices used within this study are two variations of vocabulary diversity indices. Vocabulary diversity indices are practical and lower-inference than other vocabulary indices, even within this low- inference approach to language assessment, because they don't require any decision-making about what counts as a mature word (Read, 2000). They also have a long history of use within language assessment (Lee & Muncie, 2006; Malvern & Richards, 2002; Miller et al., 2006; Olinghouse & Leaird, 2009; Olinghouse & Wilson, 2013; Simon-Cereijido & Guiterrez-Clellen, 2009; Silverman & Ratner, 2002; Uccelli & Páez, 2009; Watkins, Kelly, Harbers, & Hollis, 1995; Wood et al., 2018). For instance, studies have supported vocabulary diversity indices as measures of language disorders (Malvern & Richards, 2002; Miller et al., 2006; Morris & Crump, 1982), as measures of vocabulary to better understand cross-linguistic relations amongst various language components (Simon-Cereijido & Guiterrez-Clellen, 2009), and as indicators of English or Spanish language

ability in studies of pre-school aged emergent bilinguals (Silverman & Ratner, 2002; Uccelli & Páez, 2009; Watkins et al., 1995; Wood et al., 2018).

Corrected Type Token Ratio. Corrected Type Token Ratio (CTTR; Carrol, 1964), is a modification to the original and perhaps most common measure of vocabulary diversity – the Type Token Ratio (TTR; Klee, 1992; Malvern & Richards, 2002; Olinghouse & Leaird, 2009; Scott, 2009; Silverman & Ratner, 2002). TTR is a ratio-based measure that calculates the proportion of different words (types) to total words (tokens) in students' writing (Olinghouse & Leaird, 2009). There are also many other variants of TTR, such as Malvern & Richard's D (2002) and the Measure of Textual Lexical Diversity (MTLD; Malvern & Richards, 2002; Olinghouse & Leaird, 2009; Olinghouse & Wilson, 2013). All of these indices are considered production-independent; they are designed to control for the amount of language produced. The Corrected Type Token Ratio (CTTR) was selected because Olinghouse & Leaird (2009) found promising results for this variation of TTR and because it only requires simple statistical transformation of TTR (number of different words divided by two times the square root of total words), and not any language software that some other indices require (Olinghouse & Leaird, 2009).

Number of Different Words (NDW). Most recent studies of written vocabulary diversity indices have used CTTR or other production-independent indices (Graham, Hebert, Paige Sandbank, & Harris, 2014; Olinghouse & Leaird, 2009; Olinghouse & Wilson, 2013). However, this study will also explore a production-dependent vocabulary diversity index: Number of Different Words (NDW). NDW is simply the number of different words written (total minus repeated words); in other words, it is the type portion

of the type-token ratio (Scott, 2009). Written CTTR and NDW are similar in that they both measure the range of vocabulary employed in writing. However, they were both chosen for this study because both production-dependent and production-independent approaches are commonly used within language assessment (Keller-Margulis et al., 2016; Scott, 2009; Woolpert, 2016; Yu, 2010). According to Woolpert (2016), productivity, complexity, and accuracy are three important dimensions of written language. NDW and CTTR likely tap into these dimensions to different extents, with NDW more closely tied to the productive dimension and CTTR more closely tied to the complexity dimension (Keller-Margulis et al., 2016; Uchikoshi, Yang, Lohr, & Leung, 2016; Woolpert, 2016; Yu, 2010).

Previous Research on Written Vocabulary Diversity Indices

Though much of the research on vocabulary diversity indices has been conducted within speaking tasks, several studies have investigated vocabulary diversity indices within writing tasks as well (Olinghouse & Leaird, 2009). These studies show initial evidence that written vocabulary diversity indices are a) related to broader language outcomes and b) more promising indicators than other types of written vocabulary indices. For example, Grobe (1981) compared a long list of writing indices to determine which could best predict teacher-rated writing quality for 5th, 8th, and 11th graders. She included measures of a variety of language dimensions, including composition length, spelling, grammar, and vocabulary, and multiple measures of each construct. For vocabulary measures, she used several types of vocabulary diversity indices, including NDW and TTR, as well as other metrics (e.g., MTLD, word repeat rate, and a ratio of repeat rate to total words). Other types of vocabulary measures included word size,

percent of frequently used words, and readability scores. One important finding from Grobe (1981)'s study was that the vocabulary measures explained sizeable unique variance above and beyond the measures of production, grammar, and spelling. Furthermore, she found that the vocabulary diversity measures were the best predictors of writing quality, along with the spelling measure. Interestingly, she also argued that the predictive power of the total number of words was due to covariance with NDW, and that NDW was the more important measure.

In another study, Morris & Crump (1982) investigated whether several different grammar and vocabulary measures were able to indicate risk for language related learning disabilities (oral language, written language, basic reading, reading comprehension, and listening comprehension). The authors used a sample of non-EL students aged 9 to 15, and language disability status was identified by the school using an IQ discrepancy model. For the vocabulary measures, they compared CTTR and a measure of vocabulary intensity. They found CTTR, and not the measure of vocabulary intensity, to be a significant indicator: students who were typically developing received significantly higher CTTR scores than students with a language related learning disability (Morris & Crump, 1982).

Olinghouse and colleagues also found evidence for construct and criterion validity of written vocabulary diversity measures for non-EL students, and their studies (unlike the older studies) also reported correlation results. For example, Olinghouse & Leaird (2009) were interested in the extent to which various vocabulary measures were related to overall quality of narrative writing for second and fourth graders. They used an experimental measure in which students were given one of three randomly assigned

picture prompts and had five minutes to plan and 15 minutes to write. Research assistants scored the experimental writing samples for four types of vocabulary measures (vocabulary diversity, less frequent vocabulary, mean syllable length, and number of polysyllabic words), spelling (number of words spelled correctly), and composition length (total number of words). They also scored the writing samples for overall writing quality with an experimenter-developed rubric, which they used as their first criterion measure. Students were also administered the story construction subtest of Test of Written Language-3rd edition (TOWL-3) as a second criterion measure. Vocabulary diversity (CTTR) demonstrated moderate to strong correlations with both outcome measures. For instance, for second graders, the correlation with narrative quality on the experimental writing assessment was $r = .72$, and the correlation with the story construction subtest of the TOWL-3 was $r = .52$. In fact, vocabulary diversity was the single best unique predictor of the narrative writing quality criterion. In fourth grade, vocabulary diversity remained an important predictor, but compositional length was the strongest predictor of writing quality (Olinghouse & Leaird, 2009).

Olinghouse & Wilson (2013) similarly found that vocabulary diversity (MTLD) had a significant, though smaller relation ($r = .32, p < .01$) with narrative writing quality using the TOWL-3 story writing subtest. Vocabulary diversity was a unique predictor of story writing quality, explaining 8.4% of unique variance. Interestingly, the authors also found that vocabulary diversity did not have a statistically significant relation with quality of informative writing. This suggests that vocabulary diversity may serve as a better indicator of writing quality for some genres than others. These four studies from Grobe (1981), Morris & Crump (1982), Olinghouse & Leaird (2009), and Olinghouse &

Wilson (2013), which build on a broader base of literature on oral vocabulary diversity indices, demonstrate significant promise for written vocabulary diversity measures. However, many gaps remain to be investigated.

Important Gaps within Previous Research on Vocabulary Diversity Indices

Reliability Gaps. The first important gap to explore is the extent to which NDW and CTTR can provide reliable results in the CBM-W format. Relatively little research has been conducted on the reliability of vocabulary diversity indices in writing. First, most studies have used language software to calculate vocabulary diversity and thus more research on interrater reliability is needed for hand scored assessments, which are more feasible within schools. However, one study showed significant promise in this area. Graham et al. (2014) tested the reliability of many CBM-W indices, including CTTR. The researchers found an interrater reliability of .98 for CTTR in the second-grade sample; although in this study, papers were typed first and students were not given a time limit on their writing. Graham et al. (2014) did not report if ELs were included in their study.

For test-retest and alternate form reliability of vocabulary diversity indices, research is also limited. Results from Olinghouse & Leaird (2009) and Graham et al. (2014), however, suggest promise for the reliability of vocabulary diversity indices. Olinghouse & Leaird (2009) found that their measure of vocabulary diversity (CTTR) was the only vocabulary index explored that remained stable across two different prompts (paired *t* test revealed insignificant differences). Moreover, Graham et al. (2014) also found that written vocabulary diversity (CTTR) was the only writing measure in their study that had a generalizability coefficient that surpassed a criterion of .8, which is the

criterion typically recommended for individual screening decisions (Salvia, Ysseldyke, & Witmer, 2017). More research is needed to build off this initial evidence and examine the reliability of written vocabulary diversity measures under typical CBM-W conditions, especially with English learners and with the NDW index.

Criterion Validity Gaps. Another area in need of more investigation is the validity of vocabulary diversity measures relative to different standardized and comprehensive criterion measures. Most studies on written vocabulary diversity have used researcher or teacher ratings of writing quality as criterion measures, either with the experimental writing probes or within subtests of standardized writing tests. Additional research should investigate criterion validity with other standardized writing tasks, scored by outside observers and with more robust (i.e., not just single subtest) criterion measures. Furthermore, an important next step is to investigate whether vocabulary diversity measures within writing can also be used as indicators of other language domains in addition to writing, given that vocabulary represents an essential component of all language domains. Only Morris & Crump (1982)'s study used a written vocabulary diversity index (CTTR) as an indicator of outcomes not specific to writing. Further validity evidence with more robust criterion measures, including language domains other than writing, would lend more support for their use as indicators of language more broadly.

Gaps in Different Classes of Vocabulary Diversity Indices. More research is also needed to learn about the relative utility of using production-dependent indices, such as NDW, and production-independent vocabulary diversity indices, such as CTTR. Several CBM-W studies with elementary school students have found that other

production-independent indices (% WSC and % CWS) are less sensitive to growth over time, but have stronger criterion-related validity coefficients than the corresponding production-dependent measures (Allen, Poch, & Lembke, 2018; Jewell and Malecki, 2005; Parker, Tindal, & Hasbrouck, 1991). Furthermore, the one known CBM-W study with a small sample of elementary school ELs found evidence that production-independent and production minus accuracy indices may be more valid indicators of writing performance on standardized tests for ELs than pure production-dependent indices. However, the evidence in this study was tenuous, because findings were not consistent across time points and the subsample of ELs included only 19 students (Keller-Margulis et al., 2016). More research is needed to compare production-dependent and production-independent versions of vocabulary diversity indices, and of writing indices more generally for English learners.

Classification Accuracy Gaps. Another area of research in need of further study is the extent to which NDW and CTTR have high classification accuracy. Classification accuracy (similar to diagnostic accuracy and screening accuracy) is the ability for measures to accurately classify students into meaningful groups. An experimental measure with high classification accuracy is able to group students in roughly the same way that a trusted criterion test does (Petscher, Kim, & Foorman, 2011; Youngstrom, 2014). It needs to have *sensitivity,* or the ability to accurately denote which students are truly at-risk (based on whether they score below a criterion test's cut score) and *specificity*, or the ability to determine which students are truly not at-risk (based on whether they would score above a criterion test's cut score; Petscher et al., 2011). When tests are able to accurately discriminate between at-risk and not at-risk students, then

educators can use them to identify which students should receive supplemental support. When tests do not have strong classification accuracy, they may identify the wrong students for supplemental support, and therefore miss students who needed support, and/or identify too many students and therefore be less effective at prioritizing resources (Petscher et al., 2011).

Receiver Operating Characteristics (ROC) analyses are a common and recommended method for exploring the classification accuracy of measures (Youngstrom, 2014). ROC analyses plot sensitivity on one axis and 1- specificity on the other axis. Along with other benefits, ROC analyses provide a summary statistic, called Area Under the Curve (AUC), that estimates the overall ability of a measure to accurately classify students into meaningful groups (Youngstrom, 2014). None of the three highlighted studies, and no other known studies, have used ROC analyses to explore the classification accuracy of written vocabulary diversity indices. However, a small number of studies have analyzed the classification accuracy of the more established CBM-W indices for elementary school students and found statistically significant results for non-ELs (Ritchey & Coker, 2013; Keller-Margulis et al. 2016). On the other hand, findings from Keller-Margulis et al. (2016) suggest that classification accuracy for CBM-W indices may be different for ELs and non-ELs. While almost all AUC statistics were significant for their subsample of Native English-speaking students, no AUC statistics were significant for their subsample of ELs. The authors acknowledged that their sample size of ELs (n = 19) was small, and that their results could have been influenced by low power. Clearly, more studies are needed that explore the classification accuracy of CBM-

W indices, especially studies for vocabulary diversity indices and with a sufficient sample size of ELs.

Gaps for English Learners in Elementary School

Notably, none of the key studies on written vocabulary diversity focused on ELs – the studies from Grobe (1981) and Morris & Crump (1982) did not report whether ELs were included, and the studies from Olinghouse & Leaird (2009) and Olinghouse & Wilson (2013) excluded all students receiving English language support. There have been many more studies for vocabulary diversity measures for ELs that have used an oral language format than have used a written language format (Miller et al., 2006; Simon-Cereijido & Gutierrez-Clellen, 2009; Uccelli & Páez, 2009; Wood et al., 2018). Few studies on CBM-writing and on written vocabulary diversity indices have focused on the unique features of multilingual language development and the assessment data needed to improve services for ELs (Campbell, Espin, & McMaster, 2013; Keller-Margulis et al., 2016; Smith & Lembke, 2020). Studies for ELs in elementary school are particularly needed. Formative language assessments can help ensure ELs receive adequate supports before difficulties become entrenched. As discussed early in the introduction, ELs who do not achieve ELP by fifth grade are highly unlikely to achieve it by eighth grade (de la Torre et al., 2019). However, questions remain about what grade level is appropriate for written vocabulary diversity measures given young EL's still-developing code-related skills and English oral language skills (Mancilla-Martinez & Lesaux, 2011; Romig et al., 2017; Smith & Lembke, 2020).

Purpose and Importance of the Present Study

The present study addresses an important research gap – the need for reliable and valid formative language assessments that inform instructional decision-making for English learners. The CBM framework is a well-established approach to developing reliable, valid, and practical formative assessments that schools can use to improve their supports for students (Deno, 1992; 2003). However, much more CBM research is needed with EL participants, and with measures of oral language skills, such as vocabulary, that are particularly important for ELs. The present study tested the promise of using CBM-Writing probes and two vocabulary diversity indices (NDW and CTTR) to serve as reliable and valid indicators of broader English language abilities. Participants were ELs, who had a home language of Spanish and who were receiving Spanish and English bilingual instruction. Participants were also in second grade; second graders are still in their primary years of elementary school, but also have had multiple years of instruction in English oral language, and code related skills. To evaluate preliminary evidence for the use of this assessment approach, the study used a state issued summative assessment of comprehensive English language proficiency that was designed for ELs and includes rigorous criterion measures across all four language domains.

Research Questions

Three research questions drove this research. The primary goal of this research was to investigate the reliability (RQ1) and validity (RQ2) of the CBM-W vocabulary diversity indices, and a secondary goal (RQ3) was to explore whether vocabulary diversity indices have adequate classification accuracy characteristics that would lend further initial support for their use in screening.

Research Question One

What are the (a) interrater and (b) one-month delayed alternate form reliability correlations for the CBM-W Vocabulary Diversity Indices (NDW, CTTR)?

It was hypothesized that interrater reliability correlations would be strong, based on consistently strong findings for other CBM-W indices but that (b) alternate form reliability coefficients would likely be moderate. The hypothesis for alternate form reliability was more tenuous because results for writing measures have often been variable and a conservative one-month delayed alternate form reliability procedure was used in this study (Benson & Campbell, 2009; Gansle et al. 2002, 2006; Graham et al., 2014; McMaster & Espin, 2007). On the other hand, there have been some promising reliability results from other studies of vocabulary diversity indices (Olinghouse & Leaird, 2009; Graham et al., 2014).

Research Question Two

What are the criterion validity correlations between the October - March CBM-W vocabulary diversity scores (NDW, CTTR), and the writing, reading, speaking, and listening domain scores on an English language proficiency assessment (the English Language Proficiency Assessment for the 21st Century (ELPA21))?

It was hypothesized that both NDW and CTTR would have positive, significant correlations with the writing domain score on the ELPA21, with moderate coefficient sizes. Previous research has found small to moderate correlations between written vocabulary diversity indices (though not in the CBM-W format) and standardized assessments of writing for non-ELs (Olinghouse & Leaird, 2009; Olinghouse & Wilson, 2013). CBM-W vocabulary diversity indices and the ELPA21 writing domain tasks also share the same language domain context, writing. Therefore, for both tests, students need

not only written vocabulary knowledge, but also other skills shown to be important for writing, such as handwriting, spelling, grammar, and discourse-related skills (Abbot, Berninger, & Fayol, 2010; Juel et al., 1986; Kim & Schatschneider, 2017; Koutsoftas, 2013; McMaster & Espin, 2007; Ritchey & Coker, 2013; Romig et al., 2017). The question of whether NDW and CTTR would be related to the other three language domains was more exploratory. There is a lack of previous research investigating how these indices relate to broader language outcomes across domains. Nevertheless, small to moderate correlations between CBM-W vocabulary indices and the reading, speaking, and listening domains were also expected. This hypothesis is based on consistent research showing that vocabulary skills, measured in different ways, are theoretically and empirically related to all four language domains for monolingual and bilingual learners (e.g., Hwang et al., 2019; Koutsoftas, 2013; Kieffer, 2012; Kim, 2016; Kim & Schatschneider, 2017; Lesaux et al., 2010; Mancilla-Martinez & Lesaux, 2011; Simon-Cereijido & Gutierrez-Clellen, 2009; Staehr, 2008; WIDA Consortium, 2012). It was also reasoned that correlations between the written vocabulary diversity indices and the ELPA21 domain scores were more likely to be moderate in size for the reading and speaking domains and that correlations were more likely to be small with the listening domains, because, like writing, reading also requires written language skills and speaking also requires expressive language skills (Koutsoftas, 2013; Shanahan, 2006).

Finally, it was also hypothesized that moderate correlations were more likely between the CBM-W vocabulary diversity index scores administered later in the year (i.e., January, February, March) because they are closer in time to the ELPA21

administration and because students would have had more time to develop English language skills more broadly, and writing in particular.

Research Question Three

What are the Area Under the Curve (AUC) indices for NDW and CTTR in relation to overall classification of English language proficiency, and in relation to the individual listening, speaking, reading, and writing domain scores, on an English language proficiency assessment (the ELPA21)?

This question was also exploratory because no known studies have explored classification accuracy for written vocabulary diversity measures, in CBM-W, or in other assessment formats. Classification accuracy studies for other CBM-W indices are also limited, though some studies have shown promise for other CBM-W indices for non-ELs (Ritchey & Coker, 2013; Keller-Margulis et al., 2016). Nevertheless, based on the same reasoning outlined under the previous research question, it was hypothesized that adequate AUC indices were possible – that the CBM-W vocabulary indices may show promise in being able to identify students who need more English language support, indicated by their performance on the ELPA21.

II. METHOD

This study used data from Project Write, an investigation of bilingual writing development and assessment. The principal investigators for Project Write are Drs. Sylvia Linan-Thompson and Patrick Kennedy. Data for this project was collected throughout the 2019-2020 school year but cut short due to the COVID-19 pandemic and school closures starting midway through March.

Design

This study used descriptive, correlational, and receiver operating characteristic (ROC) curve analyses to evaluate the interrater reliability, delayed alternate form reliability, criterion validity, and classification accuracy. It used longitudinal CBM-W data from six CBM-W administrations from October-March and a criterion measure administered once between late January and March, 2020.

Participants

Second-grade students who qualified as ELs and had a home language of Spanish were eligible for the study. All students attended one large school district that partnered with Project Write to learn more about their emergent bilingual students' writing development. The participating district serves more than 40,000 K-12 students. Based on information provided by the district, about 60% of students in the district live in poverty, 16% qualify for Special education services, and 77% of students graduate high school within four years. Eight schools and eleven second-grade teachers participated in this study and were identified based on district recommendation. The eleven teachers were invited to participate and received a stipend for their participation in the study. All of the classrooms were designed to support bilingual language development in both Spanish and

English language. Nine of the eleven participating classrooms used a transitional bilingual model modeled after Dr. Kathy Escamilla's Literacy Squared model. These classrooms in the transitional model were designed for Spanish-dominant students and to provide literacy instruction primarily in Spanish. They also provided support with English language development and cross-language connections. The other two classrooms (both in one school) used a two-way immersion bilingual program. These classrooms were intended to serve a mix of Spanish-dominant and English-dominant students to facilitate peer language modeling for both languages. Instruction transitions from a ratio of 80:20 Spanish to English to 50:50 Spanish and English throughout the elementary school years.

Students

All students in the eleven classrooms were invited to participate, but only students whose parents provided written consent participated in the study. Consent forms were provided in Spanish and delivered and collected by participating teachers. A total of 242 students were included in the data file Project Write researchers provided for this study but only 175 students are included in this study. Sixty-seven students were excluded from this study because either (a) they did not qualify as ELs for the 2019-2020 school year, as determined by the district, or (b) they lacked sufficient data to contribute to analyses pertaining to the research questions of this study, due to absence or moving out of a participating classroom.

Though data from 175 students were included in this study, the number of students that were administered each CBM-W probe varied due to student absences. For administrations between October-February, the number of students administered each

probe ranged from 136-150 students. In March, there was more missing data (n = 106) because in addition to absences, one teacher did not administer the CBM-W probe before school was closed due to COVID-19.

In most cases, the district provided the researchers with information about which students qualified as ELs. However, in the five cases where data was missing for students due to late-entry into the project during the 2019-2020 school year, it was assumed that a student qualified as an EL if they were administered the ELPA21 at the end of the 2019-2020 school year. In the district, students initially qualify as ELs if their parents identify a home language other than English on a state-approved language use survey, and the student has a qualifying score on the ELPA21 Screener. After their initial placement, students take the ELPA21 annual summative assessment (the assessment used in this study) and all students who score below Proficient continue to qualify as ELs.

Student Demographics. The school district provided limited demographic information for study participants. Demographic information was missing for a small number of students; for student gender, information was missing for four students (2.3%) and for other variables, information was missing for five students (2.9%). Eighty-one (46.3%) study participants were female, 90 (51.4%) were male, 22 (12.6 %) were receiving special education services, one (0.6%) was identified as Talented and Gifted, and 21 (12%) participated in the Migrant Ed program. One-hundred percent of students identified as Hispanic or Latino, learned Spanish as a first language, and qualified as ELs for the 2019-2020 school year.

Procedures

See Table 1 for a timeline of all measures administered. The experimental CBM-W probes were delivered by all participating teachers six times throughout the second-grade school year in English, once per month from October to March. Prior to the first administration, the research team trained classroom teachers on how to administer the CBMs. To measure fidelity to the administration procedures, trained assessors from the research team observed the test administrations twice and completed assessment fidelity checklists. Fidelity to administration procedures will be compared against a criterion of 90% implemented. Completed CBMs were scored by a team of seven trained scorers. Prior to scoring independently, scorers were required to demonstrate reliability (95%) with a principal investigator. Furthermore, 20% of assessments were double-blind scored by a second scorer. The English language proficiency assessment criterion measure (ELPA21) was administered by all participating schools between late January and early March.

Table 1

Timeline of CBM-W Measures and Corresponding Prompts

Month	CBM-W Prompt
October	Write about your favorite thing to do during the summer
November	Write about what you like to do on the weekend
December	Write about what you like to do during recess
January	Write about your favorite part of the school day
February	Write about your favorite thing to do when you play inside
March	Write about your favorite season

Measures

CBM-W Vocabulary Diversity Measures

The experimental CBM measures followed typical CBM-W procedures (Romig et al., 2017). All students were given the same standardized directions and procedures in English. In alignment with recommended practice, students were given grade-level, open-ended writing prompts that were developed by the principle investigators for Project Write. CBM-W researchers and developers have used a variety of types of prompts to elicit open-ended writing, such as story starters, picture prompts, or sentence starters, and they most commonly are designed to elicit either descriptive or narrative text. In this study, the prompts asked students to write about common every-day life activities and were designed to be equally accessible to all second-grade students; they elicited descriptive writing. Example prompts included "*Write about what you like to do on the weekend" and "Write about your favorite part of the school day."* See a list of all prompts in Table 1. At each administration, teachers read aloud standardized directions and the writing prompt. They also gave students paper on which to write their response, with the prompt written at the top. After hearing the prompt, students were instructed to think about their answer to the prompt for one minute and then given five minutes to write. Previous studies on CBM-W measures have found strong interrater reliability results, with coefficients typically above .9 (McMaster & Espin, 2007), but variable test-retest and alternate form reliability results, depending on the index, study, grade, and time elapsed between administrations. Multiple studies have found reliability coefficients of .7 - .9 and above for the most common metrics (e.g., TWW, CWS, CSW), but some studies have found weaker coefficients in the .5 and .6 range for these same metrics (Gansle et al., 2002; Gansle et al., 2006; McMaster & Espin, 2007). Moderate validity statistics have also been found. Romig et al. (2017) conducted a meta-analysis on validity of well-

researched CBM-W indices in relation to various standardized writing criterion measures at different grade levels. For K-2nd graders, they found coefficients to range from $r = .46$ (WW) to $r = .58$ (Correct minus incorrect word sequences [CIWS]).

Scoring Vocabulary Diversity. Trained scorers scored the students' writing for a variety of indices related to writing and language development. The two vocabulary diversity measures, the focus of the proposed research, were derived in the following ways. First, the scorers counted the total number of words written and the number of words repetitions. A word was defined as any group of letters separated by space (Romig et al., 2017) and a repetition was defined as a word that had already been used earlier in the sample (Graham et al., 2014). Different forms of words, such as cat/cats were not counted as repetitions (Olinghouse & Leaird, 2009). The scorers then entered their scores into an online data collection survey and the research team reviewed scores for irregularities. The research team then calculated the two vocabulary diversity measures from these scores. NDW was calculated by subtracting the repetitions from the total number of words written and CTTR was calculated by dividing NDW by the square root of two times the number of total words written (Olinghouse & Leaird, 2009).

English Language Proficiency Assessment

The English Language Proficiency Assessment for the 21st Century (ELPA21)-Grade Band 2-3 is a standardized summative assessment of English language proficiency (Oregon Department of Education-Office of Teaching, Learning, and Assessment [ODE-OTLA], 2019). It is administered to all ELs in the eight states that participate in the ELPA21 consortium, which is led by Oregon; an estimated 300,000 students every year take the ELPA21 (Huang & Flores, 2018). The ELPA21 has different tests depending on

the student's grade. The second graders in the proposed research were assessed with the version designed for second and third graders. The ELPA21 is aligned with the English Language Proficiency (ELP) Standards and college- and career-ready standards developed by the Council of Chief State School Officers (CCSSO), WestEd, and the Understanding Language Initiative of Stanford University (ODE-OTLA, 2019). These ELP standards are aligned with the Common Core State Standards in English language arts and mathematics and the Next Generation Science standards. The ELPA21 measures students' skills across all domains of English language proficiency - listening, speaking, reading, and writing (ODE-OTLA, 2019). The ELPA test is delivered online in two segments. In the first segment, listening, reading, and writing items are delivered and in the second segment, speaking items are delivered. The test utilizes a variety of response formats, including selected responses, constructed responses, and extended responses (Huang & Flores, 2018; ODE-OTLA, 2019).

Scores are provided for each of these domains, including a scaled score and an ordinal score, which range from Level 1 (Beginning) to level 5 (Advanced; ODE-OTLA, 2019). The ELPA21 does not provide an overall English language proficiency composite scale score; however, it provides an overall descriptive classification that determines each student's English language proficiency status. Students are classified as: Emerging, Progressing, or Proficient. Students who receive a Proficient score pass the test and typically no longer qualify as ELs for the following school year (ODE-OTLA, 2019). The overall ELP status is derived from a combination of the ordinal domain scores. To receive a Proficient score students had to receive only Level 4 and Level 5 scores. Students who received only Level 1 and Level 2 scores across domains were categorized as Emerging,

and students were not classified as either Proficient or Emerging, received a score of Progressing (ODE-OTLA, 2019). Though a review of the ELPA21 test from Huang & Flores (2018) found that the ELPA21 has strong content validity, they also cautioned that there is limited reliability and validity statistics available for the ELPA21. The Cronbach's alpha for this study's sample was .91 for the scale scores and .88 for the ordinal domain level scores.

III. ANALYSIS

Descriptive Statistics

To gain insight into the psychometric properties of the CBM-W vocabulary diversity indices and to determine whether Pearson's *r* correlations were appropriate for research questions one and two, descriptive statistics and graphs were analyzed for all measures, including means, standard deviations, skewness, and kurtosis statistics, as well as histograms, Q-Q plots, and boxplots. Bivariate scatterplots for all planned bivariate correlations were also visually analyzed to check for linearity.

Research Question 1: Reliability

To determine the interrater reliability of the measures, an intra-class correlation coefficient (ICC) was calculated. As stated above, approximately 20% of writing samples were double-blind scored by a second rater. This resulted in 102 NDW scores that were also scored by a secondary rater. The ICC is between a) the NDW score produced by the primary rater and b) the NDW score produced by a secondary rater. ICCs are recommended for examining interrater reliability because they reflect both the degree of correlation and the extent of agreement between ratings (Koo & Li, 2016). The ICC estimate was based on a consistency, single rater, two-way random effects model (Koo & Li, 2016). A separate calculation was not necessary for CTTR scores because both the NDW scores and CTTR scores are based on the repeated words calculation from the scorers.

For delayed alternate form reliability, Pearson's *r* correlations were estimated between the NDW scores for each of five pairs of consecutively administered alternate

forms, administered approximately one month apart. The same five correlations were conducted for CTTR scores as well, for a total of 10 alternate-form correlations.

Research Question 2: Validity

To examine the criterion validity of the CBM-W vocabulary diversity scores with the ELPA21 language domain scores, bivariate correlations (Pearson's *r*) were estimated between each of the twelve CBM-W vocabulary diversity scores (six NDW scores and six CTTR scores administered between October-March) and the four scaled scores for each language domain on the ELPA21, administered once, between late January and early March. Prior to assessing the correlations between the CBM-W vocabulary diversity indices and ELPA21 domain scores, the intercorrelations amongst CBM-W vocabulary diversity indices, and amongst the ELPA21 domain scores were also calculated.

Research Question 3: Classification Accuracy

Receiver Operating Characteristic (ROC) curve analyses were conducted to assess the classification accuracy of the CBM- vocabulary diversity indices, or their ability to categorize students into meaningful groups, using ELPA21 scores as criterion measures. The resulting Area Under the Curve (AUC) statistics were used to estimate each measure's overall classification accuracy. ROC analyses first require a decision about which cut score(s) to use on the criterion variable to determine which students can be considered "at risk" and which can be considered "not at risk." Recall that the ELPA21 provides students with an overall English language proficiency score (ELP score) of Emerging, Progressing, or Proficient. Two cut scores were explored to test the CBM-W vocabulary diversity indices' ability to classify students based on their ELP scores. First, the Proficient cut score was used; students needed to receive a score of Proficient to be

considered “not at risk.” This is a meaningful dichotomization because students who score Proficient typically no longer qualify as ELs for the following school year. Second, the Progressing cut score was explored. For this cut score, students needed to receive a score of either Progressing or Proficient to be considered “not at risk.” This is also a meaningful classification because students who score below Progressing are likely in need of enhanced support with their language development, beyond what is typical for other ELs.

ROC analyses were also conducted using each of the four language domain scores, (which are scaled scores), as criteria. For these analyses, students needed to receive a domain score at or below the 20th percentile to be considered “at-risk”, based on ELPA21 scores for this study’s sample. The purpose of these analyses was to estimate the extent to which the CBM-W vocabulary indices could accurately identify students who would score in the lowest 20% of this study’s sample on the ELPA21, for each language domain. The 20th percentile is a meaningful criterion because schools who use MTSS often provide supplemental support to about the lowest 20% of students.

Software and Approach to Reducing Type 1 Error

Analyses were conducted with the statistical analysis software SPSS Version 26 (IBM Corp., 2019). Histograms were created in R (R Development Core Team, 2011). Type one error rate was set at 5%, or $p < .05$, to align with typical practice in education sciences. In other words, when the probability that results were due to chance exceeded 5%, results were not interpreted. Each research question is explored with multiple analyses, because the CBM-W indices were measured at six times and there are multiple criterion measures. Conducting multiple comparisons increases the chance that

statistically significant results will be found in error (type one error). However, due to the exploratory nature of this study, additional corrections (e.g., Bonferroni correction for multiple comparisons) were not employed; rather an emphasis on the magnitude of the coefficients and broader patterns of results are emphasized to guide future research on CBM-W vocabulary diversity indices.

IV. RESULTS

Descriptive Statistics

Descriptive Statistics for NDW

Descriptive statistics and histograms for NDW scores, across all time points, are provided in Table 2 and Figure 2. Means for NDW on each CBM-W administration revealed that, on average, students wrote between about 11 (October) and 17 (March) different words. Though means did not increase sequentially each month, on average, students wrote a greater number of different words in March at the end of data collection than they did in October.

Table 2

Descriptive Statistics for NDW by Administration

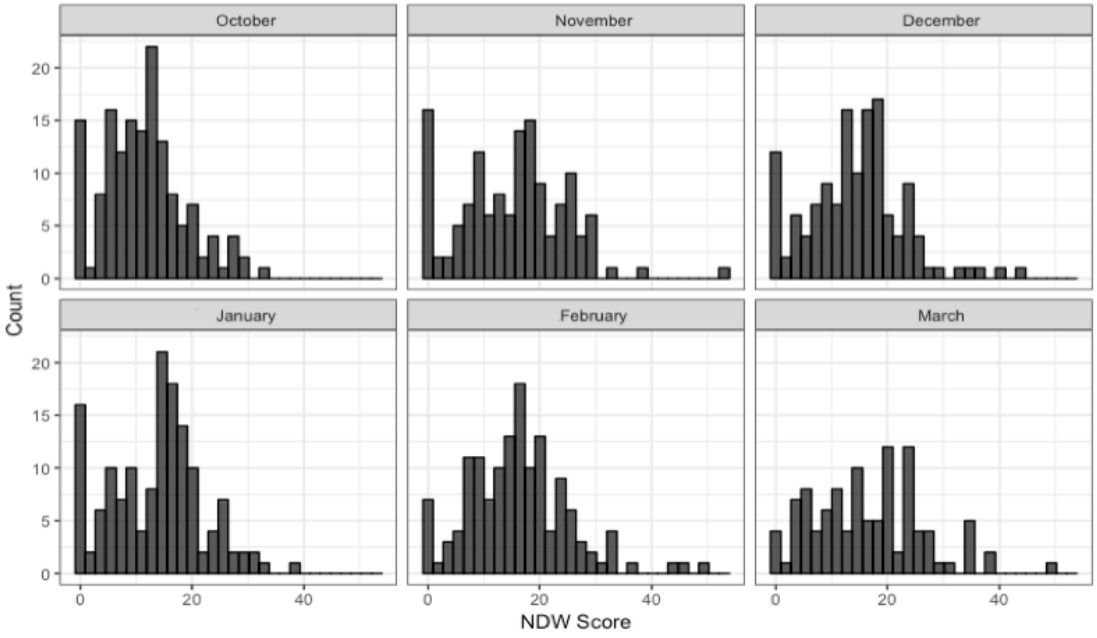
Measure	October	November	December	January	February	March
n	150.00	136.00	136.00	147.00	141.00	106.00
Mean	11.40	15.20	14.38	13.61	16.16	16.58
Median	11.00	16.00	14.50	15.00	16.00	16.00
SD	7.10	9.36	8.46	8.28	8.95	9.81
Skewness	0.50	0.35	0.45	0.13	0.76	0.52
Kurtosis	0.13	0.87	0.87	-0.23	1.58	0.32
Minimum	0.00	0.00	0.00	0.00	0.00	0.00
Maximum	32.00	53.00	44.00	39.00	50.00	50.00

The distributions of NDW scores were approximately normal for all CBM probes, based on a visual analysis of histograms (shown in Figure 2) and Q-Q plots. Skewness and kurtosis statistics were also within the generally accepted range of -2 to 2 (George & Mallery, 2010), with skew ranging from 0.13 (January) - 0.76 (February) and kurtosis ranging from -0.23 (January) - 1.58 (February). NDW scores varied widely between students. Standard deviations

ranged from about 7-9 words and the range of NDW scores varied from 0-53 words. The minimum score was zero for every probe, but maximum values ranged from 32 (October) - 53 (November). As can be seen in Figure 2, between 9-12% received a score of zero between October and January, revealing a floor effect for these first four probes. This floor effect diminished somewhat in February and March: a lower percentage of students scored zero (between 4-5%). At each administration, there were also several students (1-3) that scored significantly higher than the other students (3 SDs above the mean).

Figure 2

Histograms for NDW Scores by Administration



Descriptive Statistics for CTTR

Descriptive statistics for CTTR scores, across all time points, are provided in Table 3. Mean CTTR scores ranged from 1.99 (October) – 2.43 (March). Like NDW, CTTR scores also increased from the first administration, with a similar pattern of November- January not conforming to a month-by-month increase. The range was

narrower for CTTR scores than for NDW scores because CTTR is a ratio of different words to total words; standard deviations ranged from 0.82-1.01 and the range of scores varied from 0-4.76, across all probes. The same pattern of a significant percentage of students scoring 0 between October and January (9-12%), and a smaller percentage in February and March (4-5%) also applied to CTTR scores. For CTTR, a score of zero was three standard deviations below the mean for each administration. Compared to NDW scores, there were fewer high scoring outliers (at least 3 SDs above the mean) for CTTR scores (only 1 student in November & December, and 2 students in February).

Table 3

Descriptive Statistics for CTTR by Administration

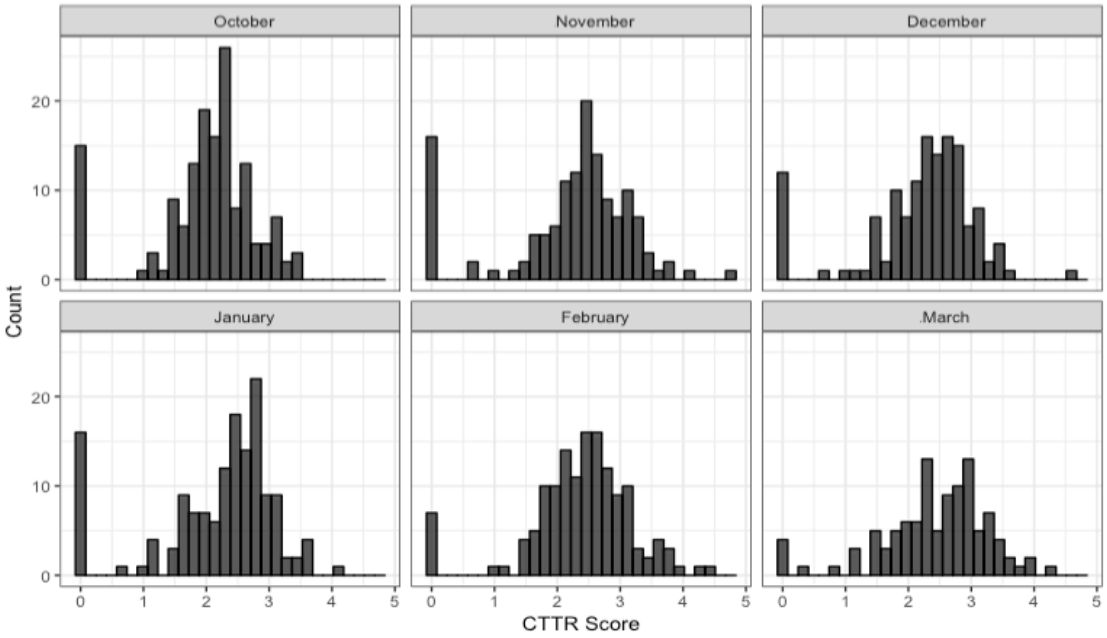
Measure	October	November	December	January	February	March
n	150.00	136.00	136.00	147.00	141.00	106.00
Mean	1.99	2.23	2.20	2.18	2.39	2.43
Median	2.12	2.44	2.37	2.41	2.46	2.57
SD	0.83	1.01	0.89	0.95	0.82	0.85
Skewness	-1.07	-0.94	-1.06	-1.08	-0.85	-0.86
Kurtosis	1.10	0.64	1.33	0.65	2.07	1.16
Minimum	0.00	0.00	0.00	0.00	0.00	0.00
Maximum	3.43	4.76	4.59	4.11	4.41	4.29

Histograms for the CTTR scores for all CBM-W probes, administered throughout the year, are shown in Figure 3. Distributions appeared approximately normal; however, a larger degree of deviation from normal was evident in the Q-Q plots, when compared with the Q-Q plots for NDW. Skewness and Kurtosis statistics were also generally higher for CTTR than for NDW, though they mostly fell within the accepted range, with the exception of the Kurtosis statistic for the

February CBM (2.07). Skewness statistics ranged from -0.85 (February) to -1.1 (October), and Kurtosis statistics ranged from 0.64 (November) – 2.07 (February).

Figure 3

Histograms for CTTR Scores by Administration



Descriptive Statistics for ELPA21

The ELPA21 was administered to 169 students; six students were not administered the test, and one student did not have a speaking domain score. Distributions of the scaled domain scores were approximately normal. Means and standard deviations are reported in Table 4. For the domain scaled scores, the highest mean was for speaking (500), followed by listening (491), reading (486), and lastly writing (477). For the domain level scores (range = 1-5), the highest mean was for listening (3.2), followed by speaking (2.30), reading (2.29), and writing (2.17). The mode for all domain level scores was 1 (39-44% of students), except for speaking, which had a mode of 3 (54% of students).

For the overall English language proficiency score, the vast majority (77%) of students who took the ELPA21 received a score of Progressing; 13% received the score of Emerging, and only 10% received a score of Proficient. This distribution is not surprising given that a much greater combination of domain level scores could result in a Progressing score. Recall that to receive an Emerging score, students needed to receive only Level 1 and Level 2 scores across domains, and to receive a Proficient score students had to receive only Level 4 and Level 5 scores; whereas, all remaining combinations of scores could result in an overall Progressing score. Moreover, the fact that the majority of the second-grade students in this sample received a Progressing score aligns with the typical developmental progression of English language proficiency, which appears to take about 4-7 years (e.g., Hakuta et al., 2000). Though not surprising, this constrained distribution of overall ELP scores was a limitation for analyzing classification accuracy based on these scores.

Table 4

Descriptive Statistics for ELPA21 Domain Scores

Domain	Mean	SD	Minimum	20th percentile	Maximum
Writing	476.75	64.91	331	422	610
Reading	485.84	61.09	345	434	609
Speaking	499.63	69.50	279	442	643
Listening	491.33	52.05	351	450	609

Research Question 1: Reliability

Inter-rater reliability was strong for the CBM-W vocabulary diversity indices, as anticipated. The ICC was .96 (CI: .95 - .97). Correlations amongst all NDW scores and CTTR scores are reported in Table 5. Correlations between NDW and CTTR were very

strongly related when taken from the same probe and time point, ranging from r = .89 - .93. Correlations amongst NDW scores, administered throughout the year and with different CBM-W probes, ranged from r = .50 - .72, and correlations amongst CTTR scores ranged from r = .42 - .65.

The relations between probes administered one month apart (delayed alternate form reliability) are bolded within Table 5. In general, alternate form coefficient sizes aligned with the hypothesis: they were consistently moderate in magnitude. For NDW, correlation coefficients were .62, .69, .69, .71, and .72, in order from smallest to largest. The smallest (.62) alternate form correlation was between January and February and stands out from the other coefficients, which ranged between .69-72. This pattern also coincides with a pattern in which the mean NDW score decreased slightly between December and January, and then increased quite notably between January and February. A possible explanation is that students lost skills over winter break, but then rebounded after renewed exposure to school for a month. This growth between January and February could possibly be causing the scores between these two scores to be more dissimilar. On the other hand, unreliability between probes is another possible explanation. The mean NDW score also increased more substantially between October and November, and the reliability between these two probes was .69.

Alternate form reliability for CTTR scores were markedly lower when compared to the reliability coefficients for NDW scores. Correlations were r = .51, .56, .58, .64, .65.

Table 5

Inter-correlations for NDW and CTTR

	October	November	December	January	February	March
October	-	**.69****	.60**	.50**	.54**	.56**
November	**.58****	-	**.71****	.58**	.69**	.53**
December	.51**	**.56****	-	**.69****	.66*	.54
January	.42**	.47**	**.64****	-	**.62****	.56**
February	.44**	.42**	.63**	**.51****	-	**.72****
March	.61**	.51**	.55**	.50**	**.65****	-

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

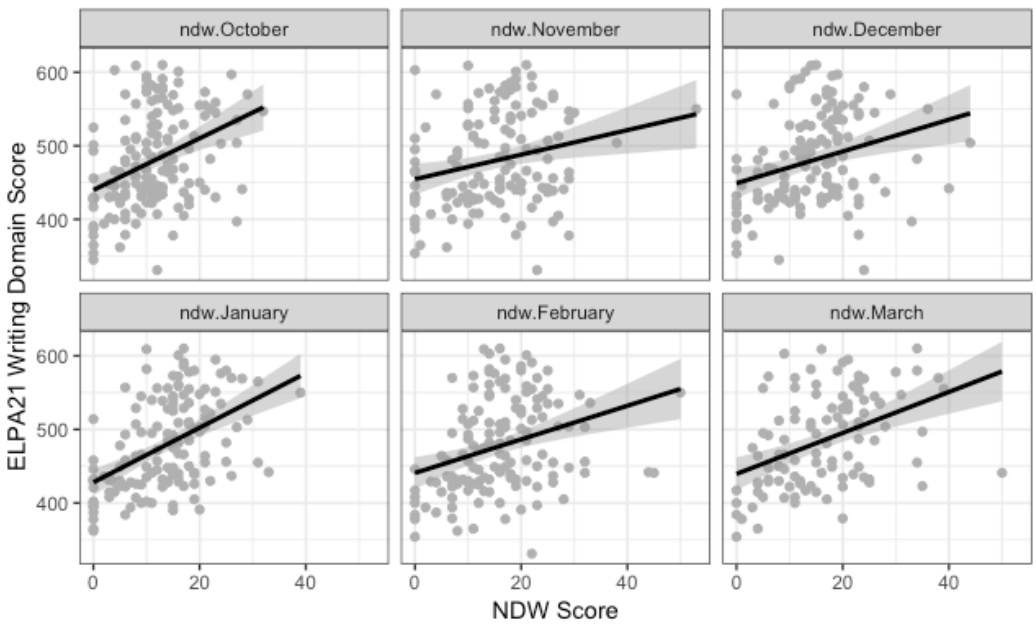
Note: NDW intercorrelations are shown above the diagonal and CTTR intercorrelations are shown below the diagonal; alternate-form correlations are bolded.

Research Question 2: Validity

Bivariate scatterplots between the CBM-W vocabulary diversity scores (NDW and CTTR) and the ELPA21 domain scaled scores indicated that an assumption of linearity was tenable. Figure 4 shows the bivariate scatterplots between NDW index scores (by administration) and the writing domain scores on the ELPA21. The black line is a fitted linear regression line (linear model) with a 95% confidence interval band. Bivariate scatterplots for all other combinations of variables showed similar patterns.

Figure 4

Bivariate Scatterplots: NDW and ELPA21 Writing, by Administration



Inter-correlations Amongst the Criterion Measures

Inter-correlations amongst the ELPA21 domain scores are reported in Table 6. These correlations were strong, ranging from r = .63 - .95. More specifically, the correlations were: r = .63 (speaking-writing), .63 (speaking-reading), .64 (speaking-listening), .67 (writing-listening), .73 (reading-listening), and .95 (reading-writing).

These correlations suggest that, in general, all language domain scores were highly inter-related, but that some language domains are more closely related to different degrees. A strong correlation between reading and writing was expected, given that they both require code-related skills and previous studies have also found strong relations between these domains; however, the size of the correlation (.95) was higher than expected because writing and reading are distinct activities, and previous studies correlating the reading and writing domains of achievement tests have reported lower

correlations (Center for Applied Linguistics, 2018; Jenkins, Johnson, & Hileman, 2004; Shanahan, 1984). For example, the technical manual for ACCESS for ELLs, another English language proficiency test, reported a correlation of $r = .64$ between reading and writing scale scores for their standardization sample (Center for Applied Linguistics, 2018, p. 93). It was also unexpected that the relation between speaking and writing *(r* = .63) was lower than the correlation between listening and writing *(r* = .67) given that both speaking and writing are expressive forms of language, but listening is not.

Table 6

Inter-correlations for ELPA21

	Writing	Reading	Speaking	Listening
Writing	-	**.95*** **	.63***	.67***
Reading		-	**.63*** **	.73***
Speaking			-	**.64*** **
Listening				-

*** *p* <.001

Correlations with the Writing Domain

The correlation coefficients between all NDW and CTTR scores and the ELPA21 language domain scores are reported in Tables 7 and 8 respectively. The correlation coefficients between all NDW scores throughout the year and the ELPA21 writing domain score ranged between $r = .24$ to $.48$ ($M = .36$). The correlation means for each language domain were calculated by standardizing the coefficients with Fisher's Z (Lenhard & Lenhard, 2014). Correlation coefficients between the CTTR scores and the language domain scores were very similar to the coefficients between NDW scores and the language domain scores. The correlation coefficients between all CTTR scores and writing were very similar to the coefficients for NDW, ranging between $r = .20$ - $.49$ (M

= .37). The results for CTTR were more variable across probes, with CTTR having generally lower correlation coefficients than NDW in October and November, and slightly higher correlation coefficients between December and March.

Table 7

Validity Correlations for NDW

ELPA Domain	October	November	December	January	February	March
Writing	.38***	.24**	.29**	.48***	.31***	.43***
Reading	.37***	.25**	.29**	.46***	.32***	.43***
Speaking	.32***	.21*	.22*	.29**	.27**	.44***
Listening	.37***	.19*	.22*	.26**	.23**	.33**

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

Table 8

Validity Correlations for CTTR

ELPA Domain	October	November	December	January	February	March
Writing	.37***	.20*	.34***	.49***	.35***	.46***
Reading	.36***	.21*	.33***	.47***	.37***	.46***
Speaking	.30***	.17	.23*	.25**	.30***	.46***
Listening	.36***	.18*	.26**	.30***	.29**	.37***

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

According to Cohen's (1992) guidelines, correlations up to .29 can be considered small, correlations between .30 and .49 can be considered medium, and correlations above .50 can be considered large. Using these guidelines, both NDW and CTTR scores can be considered to be weakly to moderately correlated with a comprehensive English written language criterion, depending on the month and probe. On average, however, these criterion validity correlations could be considered moderate.

Correlations with the Reading, Speaking, and Listening Domains

The correlation coefficients between all NDW scores throughout the year and the other language domain criterion scores on the ELPA21 ranged between r = .25 - .46 (M = .35) for reading; r = .21 -.44 (M = .29) for speaking, and r = .19 - .37 (M = .26) for listening. Results for CTTR were similar, but slightly higher for the reading and listening domains. Correlations ranged from r = .21 - .47 (M = .37) for reading; r = .17 - .46 (M = .28) for speaking; and r = .18 - .37 (M = .29) for listening.

It was hypothesized that validity results for the CBM-W vocabulary diversity indices were more likely to be moderate when using the reading and speaking domain scores as criteria, and small when using the listening domain score as a criterion measure. Results aligned with this hypothesis in part but not fully. The following comparisons between correlations are based on patterns but were not tested statistically and therefore should be interpreted with caution. In general, the patterns of correlations suggest that the relations between both NDW and CTTR and the reading domain were mostly moderate and similar but slightly lower than the correlations with the writing criterion. The relations with the speaking and listening domains were mostly small. Though the *strongest* correlations between the NDW and speaking (r = .44, in March) and CTTR and speaking (r = .46, in March) were very similar in magnitude to the strongest correlations with the reading and writing domains (r = .48 for NDW in March and .46 for CTTR in March), the *mean* correlation coefficients with the speaking criterion were small (r = .29 for NDW and .28 for CCTR). The mean correlation coefficients with the speaking criterion were more similar to the mean correlation coefficients with the listening domain (r = .26 for NDW and .29 for CTTR) than they

were with the reading (.35 for NDW, .37 for CTTR) and writing (r = .36 for NDW, .37 for CTTR) domains.

Timing and Form Differences

It was anticipated that correlations with the ELPA21 language domains scores would be strongest with the NDW and CTTR scores when administered in the second half of the year, because the ELPA was administered between late January and early March, and it was hypothesized that the CBM-W open-ended task was perhaps most appropriate after the second- grade ELs had had more time in the school year to develop their writing and English language skills more broadly. Results aligned with this expected pattern for the most part, but not fully. The second half of the year probes in January, February, and March did consistently outperform the earlier probes in November and December. However, the results for October did not align with this pattern. For the reading and writing domains, October mostly outperformed February and for speaking and listening, October consistently outperformed February, and even January and March at times.

It is also important to recognize that correlation differences amongst the CBM-W probes cannot be due to timing alone because different probes were also used for each administration, and reliability between probes was only moderate. Probes may have systematically varied in how well they indicated broader language abilities. For instance, the November and December probes consistently had the weakest validity coefficients with all language domains, for both NDW and CTTR. A reexamination of the probes used in this study also revealed that the November and December probes used a different language structure than the other probes. They asked students to write about what they

like to do (on the weekend, at recess), whereas all other probes asked students to write about their *favorite*… (thing to do in the summer, season). It is possible that the language structure used for the November and December probes elicited language that was less indicative of students' broader English language abilities than the language structure used for the other probes, as measured by performance on the ELPA21.

Research Question 3: Classification Accuracy

Results of the ROC analyses are reported in Table 9 for NDW and Table 10 for CTTR. Confidence intervals (95%) are reported in brackets.

Classification Based on Overall ELP Scores

First, ROC analyses were conducted to explore the indices' ability to accurately classify students based on their overall ELP scores, using two different ELPA21 cut scores. AUC values of .50 or below indicated that NDW or CTTR scores were no better than chance at classifying students accurately. Commonly used guidelines (Swets, Dawes, & Monahan, 2000) for interpreting AUC values suggest that AUC values > 0.70 are poor, ≥ 0.70 are fair, ≥ 0.80 good, and ≥ 0.90 are excellent, although it is worth noting that these guidelines have been critiqued for being too stringent given the imperfect reliabilities of criterion measures available (Youngstrom, 2014). When predicting to the proficiency cut score threshold, AUC values in January were statistically significant and fair in magnitude, for both NDW (.71, $p < .001$) and CTTR (.70, $p <$. 05 for CTTR). AUC values for all other CBM-W probes were insignificant when using the Proficient cut score. When using the Progressing cut score, AUC values were significant for NDW in February and March (.70, $p < .01$; .66, $p < .05$) and for CTTR in January and February

(.68, $p < .05$; .69, $p < .01$), but only the coefficient for NDW in February could be considered fair in magnitude. In sum, although three time points showed statistically significant and fair-sized AUC values, the overall pattern of results across the 24 ROC analyses conducted did not show evidence of consistently good classification accuracy for NDW and CTTR when using overall ELP scores as criteria, with either the Proficient or Progressing cut scores.

Table 9

Area Under the Curve (AUC) for NDW Receiver Operating Characteristic Curve Analyses

	Proficient	Progressing	Writing	Reading	Speaking	Listening
October	.60	.63	.70**	.69**	.64*	.66*
	[.48-.71]	[.49-.78]	[.58-.82]	[.57-.81]	[.52-.77]	[.54-.77]
November	.62	.62	.60	.67**	.61	.62
	[.48-.75]	[.48-.77]	[.46-.74]	[.54-.79]	[.47-.76]	[.50-.74]
December	.59	.61	.68*	.63*	.60	.56
	[.45-.73]	[.46-.75]	[.52-.84]	[.49-.78]	[.46-.73]	[.42-.70]
January	.71**	.64	.81***	.74***	.58	.58
	[.59-.83]	[.50-.79]	[.71-.90]	[.63-.84]	[.45-.70]	[.45-.71]
February	.60	.70**	.74***	.72***	.63*	.60
	[.50-.70]	[.56-.83]	[.62-.86]	[.60-.84]	[.50-.76]	[.47-.74]
March	.62	.66*	.81***	.73**	.71**	.65*
	[.46-.79]	[.51-.81]	[.71-.92]	[.61-.86]	[.58-.84]	[.51-.79]

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

Table 10

Area Under the Curve (AUC) for CTTR Receiver Operating Characteristic Curve Analyses

	Proficient	Progressing	Writing	Reading	Speaking	Listening
October	.52	.61	.67**	.66**	.60	.62*
	[.40-.63]	[.46-.76]	[.54-.79]	[.54-.78]	[.47-.73]	[.50-.75]
November	.59	.58	.55	.62	.58	.59
	[.45-.73}	[.42-.70]	[.40-.70]	[.49-.75]	[.43-.73]	[.46-.71]
December	.58	.60	.65*	.65*	.58	.56
	[.44-.73]	[.45-.74]	[.48-.81]	[.51-.79]	[.45-.72]	[.43-.70]
January	.70*	.68*	.79***	.73***	.59	.61
	[.58-.82]	[.55-.80]	[.70-.89]	[.63-.84]	[.47-.70]	[.48-.73]
February	.62	.69**	.73***	.70***	.62*	.61
	[.52-.72]	[.54-.84]	[.60-.86]	[.58-.82]	[.58-.82]	[.47-.74]
March	.63	.64	.80***	.70**	.67**	.62*
	[.46-.80]	[.48-.81]	[.69-.91]	[.57-.83]	[.53-.81]	[.47-.77]

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

Classification Based on Domain Scores

Next, ROC analyses were conducted to explore the indices' ability to accurately classify students based on their performance for each language domain separately, using the 20th percentile of the sample as a cut score. For these analyses, the question was: how well do the CBM-W vocabulary indices identify the students who score in the lowest 20% in each language domain on the ELPA21? Results for NDW are reported in Table 9 and results for CTTR are reported in Table 10. Results for NDW were consistently stronger than CTTR. The following descriptions will focus on NDW results.

NDW mostly served as a fair or good classifier of students who most need support in writing, the language domain most closely connected to the measures, with AUC values ranging between .60-.81 (M = .72). Consistent with the weak validity coefficients

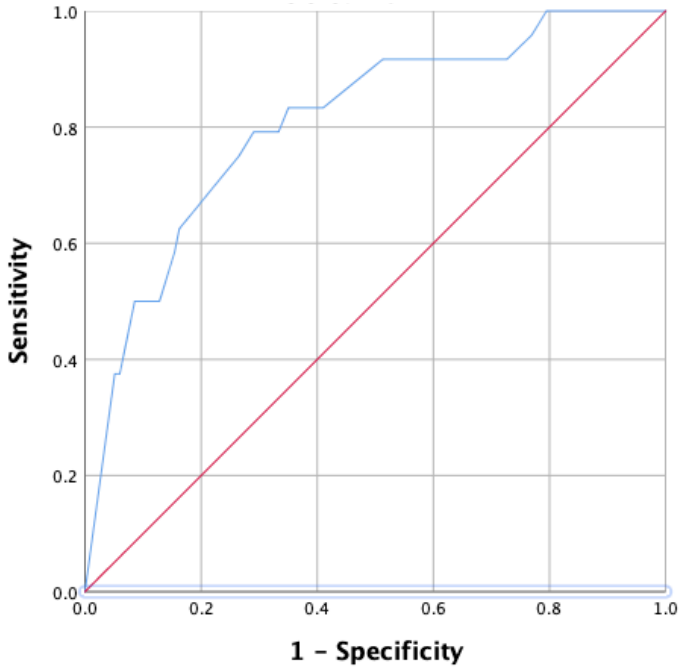
in November and December, AUC values were also poor for these November and December probes. For all other probes, AUC values could be considered either fair (≥ .70) or good (≥ .80). In October, the AUC was .70 ($p < .01$), in January it was .81 ($p < .001$), in February it was .74 ($p < .001$), and in March it was .81 ($p < .001$).

The ROC curve for NDW scores in January, which had the highest AUC value (tied with March) is shown in Figure 5. ROC curves plot sensitivity and specificity for different cut points for the screening test. ROC curves that approach the top left corner indicate better classification accuracy whereas ROC curves closer to a diagonal 45-degree line are less accurate and perform close to chance. As seen in Figure 5, the ROC curve for NDW in January (the line in blue), is distinct from the diagonal line in red, and approaches the top left corner. However, it is also evident that none of the cut scores would have both very high sensitivity (>.9) and high specificity (<.2 false positive rate). When sensitivity was at least .90 (.92), specificity was only .49 (cut score = 12.5).

NDW scores from the second half of the year also served as fair indicators of reading risk. AUC values were significant but poor in October-December (.63-.69), but they were fair from January-March (.72-.74). Results showed consistently weaker classification accuracy for NDW as a risk indicator for the oral language domains. However, it appeared to perform slightly better for speaking (.58-.71, mean = .66) than listening (.56-.66, mean = .60) and one AUC value was fair for speaking, in March.

Figure 5

ROC Curve for NDW in January (AUC = .81)



IV. DISCUSSION

In order to address gaps in research on formative language assessment for elementary school-aged English learners, this study explored two variations of vocabulary diversity indices within a writing CBM: Number of Different Words (NDW) and Corrected Type Token Ratio (CTTR). One-hundred and seventy-five second-grade Spanish-English ELs receiving bilingual instruction in both Spanish and English participated in this study. Participants completed writing CBM probes on six occasions, once per month from October to March. Descriptive statistics, reliability, validity, and classification accuracy for the two written vocabulary indices were explored and a state-issued summative assessment of English language proficiency, administered between late January and mid-March, was used as the criterion measure.

In the discussion section of this paper, I will first compare results for the production-dependent NDW index and production-independent CTTR index across research questions. Next, I will review and interpret preliminary findings on the descriptive statistics, reliability, validity, and classification accuracy for the CBM-W vocabulary diversity assessment approach, primarily focusing on the NDW index. Limitations and suggestions for future research will be embedded throughout the discussion and discussed at the end. The discussion will finish with final conclusions.

Comparison of Vocabulary Diversity Indices: NDW and CTTR

The three research questions in this study examined the reliability, criterion validity, and classification accuracy of both NDW and CTTR indices. The results showed first, that alternate form reliability coefficients were consistently higher for NDW than CTTR. This is an important finding because reliability is a fundamental pre-requisite of

measurement, and alternate form reliability is particularly important for repeated assessment for benchmark screening and progress monitoring in schools (Deno, 1992). Next, CTTR and NDW were found to have similar criterion validity with language proficiency criteria, though the results for CTTR were more variable and the means across probes were slightly higher in general. Finally, the results of research question three found that NDW had consistently higher AUC values in indexing which students would score in the bottom 20% of the sample on each of the language domain scores on the ELPA21. When contrasting the potential utility of NDW versus CTTR it is also important to consider that NDW is a much more easily understood metric compared to CTTR and has greater capacity to monitor growth over time (Tindall & Parker, 1989). The findings comparing NDW and CTTR are novel because no known other CBM-W studies have contrasted production dependent and independent indices of vocabulary indices. The preliminary evidence from this study suggest that NDW may have more promise as a written vocabulary index for ELs in second grade. Though validity coefficients were mostly higher for CTTR than NDW by a small margin, both the alternate form reliability results and classification accuracy results were stronger for NDW. These results are interesting because there seems to be increased interest in production independent indices within CBM-W research, as researchers attempt to find new ways of assessing writing quality and not just quantity (Allen et al., 2018; Keller-Margulis et al., 2016). On the other hand, is not surprising that the NDW index displayed better classification accuracy for ELs in this study. Studies have quite consistently shown that the quantity of words used in speaking and writing can help differentiate students with low and high first and second language abilities (Espin et al., 2000; Klee, 1992;

Morris & Crump; Muñoz, Gillam, Peña, & Gulley-Faehnle, 2003; Wolpert, 2016). Therefore, it is logical that an index that does not control for writing quantity would perform better in classifying students as at-risk on a measure of English written language outcomes. Furthermore, indices that are more closely aligned with the productivity dimension rather than the complexity dimension of language may be more suitable for ELs in second grade who are relatively new to English and writing.

Because of the practical advantages of NDW and preliminary findings that preference NDW, the remaining sections of the discussion will focus on results for the NDW vocabulary index. However, much more research is needed to continue to investigate the technical adequacy for these measures and extent to which they assess similar or different constructs relevant to vocabulary and English language development more broadly. For instance, it should be asked how these measures relate to other measures of vocabulary (e.g., a multiple-choice vocabulary test), other broader language outcomes (e.g., expository reading or writing), and for different populations (e.g., more advanced ELs). Future research should also ask whether these different indices explain unique variance in predicting broader language outcomes.

Descriptive Statistics

Descriptive statistics revealed that on average, students wrote between 11.40 and 16.58 different words, and that their NDW scores grew slightly throughout the year. The fact that students' scores increased, on average, across the year provides some initial support for the construct validity of NDW in the CBM-W format. If NDW is an important indicator of vocabulary and broader English language development for ELs, then it should increase throughout the year as these skills develop. Second, though an in-

depth examination of NDW growth sensitivity was beyond the scope of this paper, these preliminary results suggest that NDW could potentially be used for multiple benchmarking points throughout the year. More frequent progress monitoring may be less plausible since average growth was small and means did not increase sequentially each month. Future research should further investigate whether NDW or other measures are sensitive to growth across the year, with more advanced growth modeling techniques, and with various populations with different levels of initial skill. Measures sensitive to growth could possibly help educators monitor the general effectiveness of their instruction, whether ELs are closing English vocabulary gaps with non-ELs, or determine which students are or aren't responding to intervention.

Another promising feature of NDW as a measurement tool is that scores were distributed approximately normally. A limitation, however, was that a significant proportion of students (between 4-12%) received NDW scores of zero throughout the year. The reason for these zeros is unknown; however, one possible consideration is a lack of exposure to extended response writing activities. Some students may have been unfamiliar or unprepared for the assessment task, or it may have been too difficult at this stage in their English language and writing development. There is tentative support for this idea because there were generally smaller percentages of students scoring zero in February and March than earlier in the year. However, a visual analysis of individual scores across probes also revealed that the majority of students who scored zero on at least one probe did not score zero on multiple probes, and some even received relatively high scores on other probes. Moreover, the vast majority of students in the sample were able to successfully engage with the task. An alternative explanation, therefore, may be

that some students received zeros because they were disinterested or stumped by a particular probe, or having a difficult time focusing or complying on a given day. More research is needed to determine the cause of the zeros and to investigate whether alterations to the assessment approach could reduce them, such as incorporating more instructions on how to plan and use the one minute of think time, using different prompts, or using multiple prompts and taking the median.

Reliability

In this study, reliability of NDW scores for the students in this sample was investigated by correlating scores from two adjacent months (about 4 weeks apart), each with different probes. Coefficients ranged between r = .62-72. Salvia, Ysseldyke and Witmer (2017) provided recommendations for interpreting reliability coefficients. They suggest that coefficients of about .60 or greater are needed to make group level decisions, .70 or greater are needed for progress monitoring decisions, .80 or greater are needed for screening decisions, and .90 or greater are needed for high stakes decisions. According to these guidelines, the delayed alternate form reliability results provided evidence that NDW scores are sufficiently reliable for group-level decision making or progress monitoring decisions, but not for individual screening or high stakes decisions. According to McMaster & Espin (2007)'s review of writing CBM, these alternate form reliability results are considered moderate to moderately strong compared to other measures of writing CBM. Reliability estimates for writing tests, including more comprehensive standardized writing assessments, are generally lower than for reading tests (McMaster & Espin, 2007). Moreover, reliability estimates with a shorter time interval between alternate form administrations, would likely result in greater estimates.

For instance, alternate form reliability was $r = .91$ for the words written (WW) index, and $r = .81$ for the words spelled correct (WSC) index when administered one day apart, but $r = .64$ for WW and $r = .62$ for WSC when administered three weeks apart (McMaster & Espin, 2007).

These results are also superior to some other CBM writing measures of vocabulary. For instance, Gansle et al. (2002) found reliability estimates of $r = .01$ for number of long words and $r = .09$ for a computer scored vocabulary complexity measure of vocabulary complexity (Word Perfect). Overall, these results show that the reliability of the measures used in this study were generally in line with previous research, but also that there is significant room for improvement in measuring student's vocabulary and English language skills reliably. Additional research is needed to investigate the reliability of NDW scores across different probes and student samples, and to determine whether reliability can be improved by continuing to develop a set of equivalent probes or using statistical equating. Research can also explore the use of multiple probes at each time point to improve reliability metrics (Graham et al., 2014).

Validity and Classification Accuracy

The Writing Domain as Criterion

When using the writing domain on the ELPA21 as the criterion, which is the most proximal criterion given the CBM-W format, correlation coefficients for all NDW scores ranged between $r = .24$ and .48 ($M = .36$), depending on the administration.

Comparisons to Vocabulary Diversity Studies. These moderate correlations are similar to previous findings on written vocabulary diversity indices. The most recent studies on written vocabulary diversity indices for elementary schoolers, from

Olinghouse & Leaird (2009) and Olinghouse & Wilson (2013), also found mostly moderate correlations between vocabulary diversity indices (production independent indices CTTR and MTLD) and writing criterion scores. Olinghouse & Leaird (2009) found a correlation of $r = .52$ between a concurrent CTTR score and Story Construction subtest score of the TOWL-3, for second graders, and Olinghouse & Wilson (2013) found correlations of $r = .32$, .22, and .14 ($p > .05$) between a concurrent MTLD score and holistic researcher ratings for a narrative probe, persuasive probe, and informative probe respectively.

Though similar to previous results, the results from this study also add to the literature in several ways. First, and perhaps most importantly, this is the first known study to show moderate relations between written vocabulary diversity indices and broader writing outcomes for a sample of English learners in U.S. K-12 schools. Second, the criterion measure used in this study, the ELPA21, was much more distal and comprehensive compared to previous studies of vocabulary diversity indices. The ELPA21 writing domain score is derived from multiple types of tasks, including word building and sentence building, a descriptive constructed response, and two extended responses– one for narrative writing and one for persuasive writing. English language proficiency assessments like the ELPA21 are also highly relevant criterion measures for ELs because they were specifically designed to assess English language skills for ELs and because they are used to determine EL status and services in schools. This study is also an important extension of previous written vocabulary diversity studies because it incorporated a production-dependent index that is more interpretable, lends itself better to progress monitoring, and may also be more appropriate for ELs at this stage in their

language development. Indeed, in this study, NDW was also found to be more reliable and have stronger classification accuracy.

Lastly, these results expand upon the previous studies by including a preliminary analysis of classification accuracy. If vocabulary indices are to be used in the future for helping determine which students need more intensified English language supports, then there should be significant overlap between the students who perform low on the CBM-W measures and the students who receive low scores on important criterion measures like the ELPA21. However, this is the first known study to analyze the classification accuracy of written vocabulary indices. Results from this study indicated that AUC values (overall estimate of classification accuracy) for NDW relative to the writing domain score on the ELPA21 ranged from .60-.81 (M = .72) across probes. Most of the AUC results could be considered either fair ($\geq$ .70) or good ($\geq$ .80), according to commonly used guidelines for interpreting AUC values (Youngstrom, 2014).

Comparisons to Other Measures

Though the results in this study complement and extend previous research on written vocabulary diversity indices, they also show that NDW and CTTR are not, by themselves, *strong* indicators, or classifiers of language outcomes, especially when compared to the performance of CBM measures of reading (Romig et al., 2017; National Center on Intensive Intervention, n.d.). Yet, validity and classification accuracy for writing CBMs and standardized writing assessments are consistently lower than results for reading CBMs (Romig et al. 2017). Thus, other CBM-W indices are a necessary comparison. There are no known well-established criteria for judging CBM-W indices. However, Romig et al. (2017)'s meta-analysis, though not specific for ELs, offers the

opportunity to compare the NDW criterion validity results in this study to previous research on the most well-established CBM-W indices. Romig et al. (2017) averaged the highest correlation coefficients from each study reviewed and analyzed means specifically for students in the K-2nd grade range. The mean correlation coefficients were .37 for WW, .44 for WSC, .51 for CWS, and .60 for CIWS. In this study, the highest correlation coefficient between NDW and the ELPA21 writing domain score was r = .48 (January score). Thus, the criterion validity for NDW in this study is in line with most of the well-established CBM-W indices, but significantly lower than the CIWS index score.

Fewer studies are available for contextualizing the classification accuracy results for CBM-W indices, but Ritchey & Coker (2013)'s general population study, and Keller-Margulis et al. (2016)'s study with both non-EL and EL subsamples, can be used for a preliminary comparison. For the subsample of non-ELs, Ritchey & Coker (2013) reported AUC values between .75 (TWW) and .85 (CWS). Keller-Margulis et al. (2016) reported AUC values between .55 (WSC winter) and .90 (CIWS winter). However, for the subsample of ELs, none of the AUC values were statistically significant. Results for the subsample of ELs should nevertheless be treated with caution because the subsample of ELs in Keller-Margulis' (2016)'s study was very small (n = 19). Given the results across these studies, it appears that the NDW results in this study were comparable to the other CBM-W indices for the most part, but generally lower than the results from other studies on CWS and CIWS.

The Multi-faceted Nature of Writing

The fact that the NDW results in this study were weaker than validity coefficients for CWS and CIWS in other studies, could potentially suggest that NDW may be too simple of an index. CWS directly quantifies both spelling and grammar components, and CIWS does the same while also tapping into the complexity of the student's writing sample to a greater extent by reducing some but not all of the effects of production. The stronger results for these indices support the idea that the most predictive CBM-W indicators are those that best capture the multi-faceted nature of writing (Keller-Margulis et al., 2016; McMaster & Espin, 2007; Romig et al., 2017). Given that, it may not be reasonable to expect that a single, relatively simple, index, like NDW could be strong indicator and classifier of English written language. Vocabulary is an important component of language, but so also are skills such as spelling, handwriting, higher-order linguistic skills (e.g., grammar, discourse-level skills), and executive functioning-related skills, such as working memory, and organization of ideas (Juel et al., 1986; Kim & Schatschneider, 2016; McMaster & Espin, 2007; Ritchey & Coker, 2013; Romig et al., 2017).

On the other hand, the CBM writing measurement approach naturally requires the integration of many language subskills in ways that make it more likely that a single index score like NDW can serve as broader indicators of language outcomes. The CBM-W vocabulary indices are embedded within an authentic writing task and vocabulary is measured in an embedded, comprehensive, and context-dependent way, as advocated for in Read and Chapelle's (2001) three-dimensional framework for vocabulary assessment. The authentic writing task demands the integration of vocabulary knowledge with other important writing skills. For instance, students need to apply their English vocabulary

knowledge to a written discourse and by transcribing their ideas into written words (e.g., Babayiğit, 2014; Juel et al., 1986; Kim & Schatschneider, 2017).

At the same time, given the relatively stronger performance of measures like CWS and CIWS, it may be fruitful to investigate whether a vocabulary diversity index score that was manipulated to directly asses multiple language skills, including vocabulary, and manipulated to better capture the complexity of students' writing, could prove to be a more powerful global indicator of writing for ELs, and potentially other language domains. For instance, NDW does not penalize students for spelling errors and a new index that counts the number of correctly spelled different words could also be tested. Moreover, similar to the approach used with the CIWS index, researchers could try subtracting repeated words from the number of different words, which may potentially better assess whether a student's vocabulary use is complex, without sacrificing reliability and sensitivity to growth, as seems to be the case with CTTR, and without controlling for production (Espin et al., 2000).

Alternatively, a multivariate approach, as opposed to a single general outcome measure approach, that includes a vocabulary index such as NDW, might be equally worthy of investigation. It is possible that a composite score, made up of multiple CBM-W indices that assess distinct language components, could not only serve as a better screener of broader language outcomes, but also be more instructionally useful for educators. This multi-component approach is similar to the rubric-based multi-trait approach to evaluating writing, but would also lend itself to more objectivity, greater reliability, and sensitivity to growth by using countable analytic scores (Tindall & Parker, 1989). This multivariate approach might be more instructionally relevant and have

stronger face validity with teachers because educators would potentially be able to use the results on different measures to learn about the strengths and weaknesses of their students across different components or dimensions of writing. Indeed, Gansle et al. (2006) reported that teachers have expressed skepticism that single, simple analytic scores are indicative of writing quality and can help guide their instructional planning. One example of a potential multivariate approach would be to use both the NDW score and the CWS score to glean information about both vocabulary and spelling and determine whether a student's English writing difficulty is more likely due to a lack of lower conventional English spelling skills or productive English vocabulary knowledge. Similarly, it may be useful to know whether a student has a profile of relative strengths and weaknesses in production-independent or -dependent scores (Tindall & Parker, 1989).

Future research is needed to test different multivariate approaches to determine whether a composite of indices can produce improved validity and classification accuracy results and also provide information about distinct language constructs that could guide instruction and interventions. Moreover, evidence would be needed to show that vocabulary diversity indices contribute unique information to these models. However, this research may be especially worthwhile to better serve ELs. A focus on global indicators, which penalize students for spelling errors, could possibly over-identify ELs whose code-based skills may look poor, but actually be developmentally appropriate. Studies have shown that spelling errors that ELs make in early elementary school are often developmentally appropriate and not cause for additional intervention (Figueredo, 2006; Gort, 2006; Joy, 2011; Howard, Green, & Arteagoitia, 2012). For instance, Howard

et al. (2012) found that although second-grade Spanish-dominant bilingual learners made frequent cross-linguistic spelling errors, this pattern disappeared by fourth grade. Though there is much more research needed to support the use of CBM-W vocabulary diversity indices as indicators of the writing domain of English language proficiency for ELs, either within a general outcome measure approach or multivariate approach, this research is important. The current lack of focus on integrating vocabulary measures into formative language assessment for elementary school students is problematic for English learners, for whom English vocabulary instruction is particularly important.

Reading, Speaking, and Listening Domains as Criteria

Relatively few validation studies of CBM-W have utilized non-writing criterion-based measures. However, this study explored the CBM-W Vocabulary diversity indices' validity and classification accuracy with all four language domains of ELP and overall ELP scores, and not just the most proximal written language domain. This was an important exploration because ELs are in the process of developing comprehensive English language proficiency across all four language domains and their EL status and services are dependent on their test scores on comprehensive English language proficiency tests. Moreover, as discussed in the introduction, vocabulary is a fundamental component of all language domains, and thus it is important to explore the extent to which CBM-W vocabulary indices provide information about ELs' comprehensive language skills.

For the non-writing language domains, it was hypothesized that validity results were more likely to be moderate in relation to the reading domain, because of its written form, and in relation to the speaking domain, because of its expressive form.

Comparisons across language domains in this study are based on patterns and not statistical tests. With this caution in mind, the results seemed to support the hypothesis for reading, but not speaking. In general, the indices had slightly lower validity and classification accuracy for the reading domain, but the results were also moderate and within a similar range as for writing. In contrast, validity and classification accuracy were mostly weak in relation to the speaking and listening oral language domains. They were also weak classifiers of EL's overall English language proficiency scores on the ELPA21, which are based on performance across all four domains.

The moderate results for the reading domain provided a small degree of support for the idea that expressive language tasks can serve as indicators of students' skills in the receptive language domains. As discussed in the introduction, previous studies have found that expressive oral vocabulary measures have outperformed receptive vocabulary measures in predicting comprehensive language outcomes for ELs, including performance in the receptive language domains (Hwang et al., 2019; Kieffer, 2012). The present study did not compare expressive CBM-W indices to any receptive indices; however, the moderate validity and classification accuracy results for the receptive reading domain may provide some justification for continued research into whether writing tasks can be used within formative assessment of reading.

On the other hand, this study did not support the hypothesis that the CBM-W vocabulary diversity indices could be moderately indicative of performance on the speaking English language domain due to the expressive nature of the task. In general, correlations and AUC values were higher for speaking than for listening, but neither could be considered moderate. The moderate results for the writing and reading domains

and the weak validity and classification accuracy results for the oral language domains, is likely due, at least in part, to the fact that second graders are still in the process of mastering the decoding and encoding skills required for reading and writing but not listening and speaking (Lesaux et al., 2010; Language and Reading Research Consortium [LAARC], 2015; Shanahan, 2006; Uchikoshi et al., 2016; Wise, Sevcik, Morris, Lovett, & Wolf, 2007). As discussed in the introduction, written formats are more practical for formative assessment purposes in schools compared to oral formats, primarily because they can be group administered and leave a permanent product that can be easily scored. Moreover, there is research showing that although code-related skills are still important predictors of language outcomes in second grade, they also start to become less constraining on models of reading and writing around this time, as students develop more automaticity with the code (Catts et al., 2005; Juel et al., 1986; LAARC, 2015; Mancilla-Martinez & Lesaux, 2011). For these reasons, this study tested whether CBM-W vocabulary indices could serve as general indicators of broader English language proficiency, across all language domains. The results clearly showed, however, that the indices were relatively weak indicators of the oral language domains and overall English language proficiency, as measured by the ELPA21. For second-grade ELs similar to this sample, it may make sense to continue exploring written formats of vocabulary indices for assessing written language outcomes (perhaps both writing and reading), but oral formats of vocabulary indices for assessing the listening and speaking domains. NDW and CTTR have been explored within oral retell formats quite extensively for emergent bilingual preschoolers, and to some extent for elementary school aged students, but more research in this area is needed (Bitetti & Hammer, 2016; Miller et al., 2006; Muñoz et al.,

2003; Iglesias & Rojas, 2012; Uchikoshi et al., 2016; Wood et al., 2018). It may also be fruitful to explore whether CBM-W vocabulary diversity indices are more broadly indicative of both written and oral language outcomes in the upper elementary school grades when code-based skills and handwriting are more automatic and differences between oral and written language are less pronounced (Lesaux et al., 2010; Shanahan, 2006).

Limitations and Future Research

The present study expanded the limited research on formative language assessment for English learners. Limitations have been noted throughout the discussion; however, there are several other limitations worth noting that could drive future research. First, only English measures were explored in this study. Bilingual assessment research is direly needed and authentic language assessment approaches like CBM-W have the capacity to be used bilingually because they are not language specific (Miller et al., 2006; Thordardottir, Rothenberg, Rivard, & Naves, 2006; Uchikoshi et al., 2016). Second, only one type of CBM-W format was explored in this study. Future research could explore results for the same indices in different formats. For example, probes can be manipulated to elicit different genres of writing (e.g., narrative, persuasive, expository) or to provide more scaffolds for students when writing (e.g., pictures, word banks; Campbell et al., 2013; McMaster & Campbell, 2008; Johansson, 2009; Smith & Lembke, 2020; Yu, 2010).

Third, only one assessment was used for criterion measures. Though using a comprehensive assessment designed for English learners to assess the criterion validity of CBM-W vocabulary indices is an important new advancement, more research is needed

with other English language proficiency tests and other criterion measures. This is especially important because the technical adequacy of the criterion measure is important and limited technical adequacy evidence is publicly available for the ELPA21 (Huang & Flores, 2018; Youngstrom, 2014). Moreover, there were some unexpected intercorrelation results, such as the very strong intercorrelation (r = .95) between the reading and writing domains for this sample. Similarly, each English language proficiency test uses different methodology for determining the cut scores. The methodology affects the distribution of students receiving Emerging, Progressing, or Proficient overall ELP scores, which in turn influences the classification accuracy of the formative measures (Youngstrom, 2014). In this study, the vast majority (77%) of students who took the ELPA21 received a score of Progressing; this is a result of the true levels of English language proficiency of this sample, but also the cut score methodology. Therefore, additional studies are needed with different samples, as well as alternative criterion measures that use different cut score methodologies than used by the ELPA21.

Finally, a noteworthy limitation of this study is that all participating students came from one school district and all were learning Spanish and English; furthermore, little information is known about the students' instructional context, beyond that the schools endorsed either a transition or two-way immersion approach to bilingual instruction. Therefore, results from this study do not necessarily generalize to English learners broadly, and more research is needed to understand whether these indices would be useful for different populations and instructional contexts. Several instructional factors could theoretically influence the technical adequacy of CBM-W vocabulary indices. For example, a classroom or school's relative emphasis on the different language domains

(e.g., time spent reading vs. writing), and different language components (e.g., time spent spelling words vs. writing sentences or narratives) could influence the extent to which a writing-based vocabulary index is indicative of broader language outcomes. There is indication from the field that, in general, elementary school students are not receiving recommended amounts of writing instruction or quality vocabulary instruction (Gilbert & Graham, 2010; Juel, Biancarosa, Coker, & Deffes, 2003; Lesaux & Kieffer, 2010). Perhaps the CBM-W vocabulary indices would have increased utility within instructional contexts that place greater emphasis on quality vocabulary and writing instruction. Similarly, the ESL or bilingual model that schools use may also be influential on the relevance of a CBM-W based vocabulary measure in English, given that students' language skills progress differently depending on the model of instruction they receive (Francis, Lesaux, & August, 2006). For instance, some students may have received significant exposure to English but have only practiced writing in Spanish during Spanish literacy classes and have relatively weaker writing skills in English at that moment. These students may have a greater discrepancy in their written and oral English language skills relative to students with greater English writing exposure.

Summary and Final Conclusions

The results of this study provide initial evidence on the reliability, validity, and classification accuracy for two vocabulary indices embedded within experimental CBM writing probes: the production dependent NDW index, and the production independent CTTR index. All participants in the study were English learners and emergent bilinguals in Spanish and English. In general, results for NDW and CTTR were similar, but they also suggested more promise for NDW due to the slightly stronger reliability and

classification accuracy results, and due to the fact that NDW is easier for educators to interpret and monitor progress over time. Overall, NDW had promising measurement characteristics for this sample: distributions were approximately normal and means increased from October to March. Interrater reliability was strong and delayed alternate form reliability for NDW were mostly moderate compared to other CBM writing assessments, but not strong enough for screening decisions or individual decision making at this point. Validity and classification accuracy results were also mostly moderate when using the robust criterion of the ELPA21 writing domain score and to a similar, though slightly smaller extent, when using the ELPA21 reading domain score. On the other hand, they were mostly weak when using the oral language domains and when using overall English language proficiency as criterion measures.

The exploratory nature of this study and the moderate results were not strong enough to merit a recommendation for the use of written vocabulary diversity indices with these probes and procedures in schools at this time. However, the moderate results for writing, and to some extent reading, make an argument for further exploration of vocabulary indices such as NDW within a CBM-W format. Future research is needed to continue to develop and test new variations of this assessment approach, including with different populations, with other analytic approaches, and potentially in combination with other measures.

Formative assessment of language has become increasingly recognized as integral to identifying students that need supplemental support, evaluating instructional programming based on screening data, providing diagnostic information about which skills need targeting for individual students, and monitoring growth and response to

evidence-based interventions. However, most of this work has been conducted with non-ELs and within primarily the reading domain only (Campbell et al., 2013; Keller-Margulis et al., 2016; Lesaux & Kieffer, 2010; Smith & Lembke, 2020). Moreover, not nearly enough formative language assessment research, including within CBM-W, has focused on oral language skills, such as vocabulary, that are especially crucial for ELs (Adlof & Hogan, 2019; Biemiller, 2012; Catts et al., 2005; Pearson et al., 2012; Smith & Lembke, 2020). It is urgent that researchers continue to investigate assessment methods that can provide schools with the data needed to make informed decisions that better unlock the potential of the growing population of emergent bilingual ELs in American schools.

.

REFERENCES CITED

Abbott, R. D., Berninger, V. W., & Fayol, M. (2010). Longitudinal relationships of levels of language in writing and between writing and reading in grades 1 to 7. *Journal of Educational Psychology*, *102*(2), 281.

Adlof, S. M., & Hogan, T. P. (2019). If we don't look, we won't see: Measuring language development to inform literacy instruction. *Policy Insights from the Behavioral and Brain Sciences*, *6*(2), 210–217. https://doi.org/10.1177/2372732219839075

Albers, C. A. & Mission, P. L. (2014). Universal screening of English language learners: Language proficiency and literacy. In R. J. Kettler, T. A. Glover, C. A. Albers, C. A., & K. A. Feeney-Kettler (Eds.), *Universal screening in educational settings: Evidence-based decision making for schools*. American Psychological Association.

Allen, A. A., Poch, A. L., & Lembke, E. S. (2018). An exploration of alternative scoring methods using curriculum-based measurement in early writing. *Learning Disability Quarterly*, *41*(2), 85–99. https://doi.org/10.1177/0731948717725490

American Speech-Language-Hearing Association. (2004). Preferred practice patterns for the profession of speech-language pathology [preferred practice patterns]. Available from www.asha.org/policy.

August, D., Carlo, M., Dressler, C., & Snow, C. (2005). The critical role of vocabulary development for English language learners. *Learning Disabilities Research and Practice*, *20*(1), 50–57. https://doi.org/10.1111/j.1540-5826.2005.00120.x

Babayiğit, S. (2014). Contributions of word-level and verbal skills to written expression: Comparison of learners who speak English as a first (L1) and second language (L2). *Reading and Writing*, *27*(7), 1207-1229.

Baker, S. K., Simmons, D. C., & Kame'enui, E. J. (1995). Vocabulary acquisition: Synthesis of the research. National Center to Improve the Tools of Educators, College of Education, University of Oregon.

Beck, I. L., McKeown, M. G., & Kucan, L. (2002). *Bringing words to life: Robust vocabulary instruction*. Guilford Press.

Benson, B.J., Campbell, H.M. (2009). In Troia, G. A. (Ed.), *Instruction and assessment for struggling writers: Evidence-based practices*. Guilford Press.

Biemiller, A. (2012). Teaching vocabulary in the primary grades: Vocabulary instruction needed. In E. J. Kame'enui & J. F. Baumann (Eds.), *Vocabulary instruction: Research to practice*, *second edition*. 34-50.

Bitetti, D. & Hammer, C. S. (2016). The home literacy environment and the English narrative development of Spanish-English bilingual children. *Journal of Speech, Language, and Hearing Research*, 59, 1159–1171.

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

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

Catts, H. W., Hogan, T. P., & Adlof, S. M. (2005). Developmental changes in reading and reading disabilities. In H. Catts & A. Kamhi (Eds.), *Connections between language and reading disabilities* (pp. 25–40). Mahwah, NJ: Erlbaum.

Center for Applied Linguistics. (2018). Annual technical report for Access for ELLS 2.0 Online English Language Proficiency Test, series 401 , 2016 – 2017 administration. Retrieved from https://www.cde.state.co.us/assessment/accessforellsonlinetechreport

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155.

Council of Chief State School Officers. (2014). *English Language Proficiency Standards with Correspondences to the K-12 Practices and Common Core State Standards*. Retrieved from: https://www.oregon.gov/ode/students-and-family/equity/EngLearners/Pages/EnglishLanguageProficiencyStandards.aspx

Coyne, M. D., McCoach, D. B., Loftus, S., Zipoli Jr, R., & Kapp, S. (2009). Direct vocabulary instruction in kindergarten: Teaching for breadth versus depth. *The Elementary School Journal*, *110*(1), 1-18.

Cummings, K. D., Stoolmiller, M. L., Baker, S. K., Fien, H., & Kame'enui, E. J. (2015). Using school-level student achievement to engage in formative evaluation: comparative school-level rates of oral reading fluency growth conditioned by initial skill for second grade students. *Reading and Writing*, *28*(1), 105–130. https://doi.org/10.1007/s11145-014-9512-5

de Brey, C., Musu, L., McFarland, J., Wilkinson-Flicker, S., Diliberti, M., Zhang, A., Branstetter, C., & Wang, X. (2019). Status and trends in the education of racial and ethnic groups 2018. *National Center for Educational Statistics*, *NCES 2019-*, 181. https://doi.org/10.1037/e571522010-001

de La Torre, M., Blanchard, A., Allensworth, E. M., & Freire, S. (2019). *English learners in Chicago Public Schools: A new perspective*. https://doi.org/10.2019/PDF/jh.design@rcn.com

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

Deno, S. (1992). The nature and development of curriculum-based measurement. *Preventing School Failure: Alternative Education for Children and Youth*, *36*(2), 5-10.

Deno, S. L. (2003). Curriculum-based measures: Development and perspectives. *Assessment for Effective Intervention*, *28*(3–4), 3–12. https://doi.org/10.1177/073724770302800302

DiCerbo, P. A., Anstrom, K. A., Baker, L. L., & Rivera, C. (2014). A review of the literature on teaching academic English to English language learners. *Review of Educational Research*, *84*(3), 446–482. https://doi.org/10.3102/0034654314532695

Dressler, C., & Kamil, M. L. (2006). First-and second-language literacy. In D. E. August & T. E. Shanahan (Eds.), *Developing Literacy in Second-Language Learners*, *Report of the National Literacy Panel on Language-Minority Children and Youth* (249-267). Lawrence Erlbaum Associates Publishers.

Espin, C., Shin, J., Deno, S. L., Skare, S., Robinson, S., & Benner, B. (2000). Identifying indicators of written expression proficiency for middle school students. *The Journal of Special Education*, *34*(3), 140-153.

Figueredo, L. (2006). Using the known to chart the unknown: A review of first-language influence on the development of English-as-a-second-language spelling skill. *Reading and Writing*, *19*(8), 873-905.

Foorman, B., Beyler, N., Borradaile, K., Coyne, M., Denton, C. A., Dimino, J., Furgeson, J., Hayes, L., Henke, J., Justice, L., Keating, B., Lewis, W., Sattar, S., Streke, A., Wagner, R., & Wissel, S. (2016). Foundational skills to support reading for understanding in kindergarten through 3rd grade (NCEE 2016-4008). Washington, DC: National Center for Education Evaluation and Regional Assistance (NCEE), Institute of Education Sciences, U.S. Department of Education. Retrieved from the NCEE website: http://whatworks.ed.gov.

Francis, D.J., Lesaux, N., & August, D. (2006). Language of instruction. In D. August & T. Shanahan (Eds.), *Developing Literacy in Second-Language Learners: Report of the National Literacy Panel on Language Minority Children and Youth* (365-413). Lawrence Erlbaum Associates Publishers.

Gansle, K. A., Noell, G. H., Resetar, J. L., Williams, K. L., & VanDerHeyden, A. M. (2006). The technical adequacy of curriculum-based and rating-based measures of written expression for elementary school students. *School Psychology Review*, *35*(3), 435–450.

Gansle, K. A., Noell, G. H., VanDerHeyden, A. M., Naquin, G. M., & Slider, N. J. (2002). Moving beyond total words written: The reliability, criterion validity, and time cost of alternate measures for curriculum-based measurement in writing. *School Psychology Review*, *31*(4), 477-497.

George, D., & Mallery, M. (2010). *SPSS for Windows Step by Step: A Simple Guide and Reference, 17.0 update*. Boston: Pearson.

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

Gilbert, J., & Graham, S. (2010). Teaching writing to elementary students in grades 4–6: A national survey. *The Elementary School Journal*, *110*(4), 494-518.

Goldenberg, C., & Coleman, R. (2010). *Promoting academic achievement among English learners: A guide to the research*. Corwin Press.

Goldenberg, C., Reese, L., & Rezaei, A. (2011). Contexts for language and literacy development among dual-language learners. In A. Durgunoğlu & C. Goldenberg (Eds.), *Language and literacy development in bilingual settings* (3-25). The Guilford Press.

Gort, M. (2006). Strategic codeswitching, interliteracy, and other phenomena of emergent bilingual writing: Lessons from first grade dual language classrooms. *Journal of Early Childhood Literacy*, 6(3), 323–354. https://doi.org/10.1177/1468798406069796

Gottlieb, M. (2016). *Assessing English language learners: Bridges to educational equity: Connecting academic language proficiency to student achievement*. Corwin Press.

Graham, S., Hebert, M., Paige Sandbank, M., & Harris, K. R. (2014). Assessing the writing achievement of young struggling writers: Application of generalizability theory. *Learning Disability Quarterly*, *39*(2), 72–82. https://doi.org/10.1177/0731948714555019

Grobe, C. (1981). Syntactic maturity, mechanics, and vocabulary as predictors of quality ratings. *Research in the Teaching of English*, *15*(1), 75–85. Retrieved from http://www.jstor.org.libproxy.uoregon.edu/stable/40170871

Hakuta, K., Butler, Y., Witt, D. (2000). How Long Does It Take English Learners to Attain Proficiency? California University, Santa Barbara. Linguistic Minority Research Institution. Retrieved from: https://files.eric.ed.gov/fulltext/ED443275.pdf

Hammer, C. S., Jia, G., & Uchikoshi, Y. (2011). Language and literacy development of dual language learners growing up in the United States: A call for research. *Child Development Perspectives*, *5*(1), 4–9. https://doi.org/10.1111/j.1750-8606.2010.00140.x

Howard, E. R., Green, J. D., & Arteagoitia, I. (2012). Can yu rid guat ay rot? A Developmental Investigation of Cross-Linguistic Spelling Errors Among Spanish-English Bilingual Students. Bilingual Research Journal, 35(2), 164–178. https://doi.org/10.1080/15235882.2012.703637

Huang, B. H., & Flores, B. B. (2018). The English language proficiency assessment for the 21st century (ELPA21). *Language Assessment Quarterly*, *15*(4), 433-442.

Hwang, J. K., Mancilla-Martinez, J., Mcclain, J. B., Oh, M. H., & Flores, I. (2019). Spanish-speaking English learners' English language and literacy skills: The predictive role of conceptually scored vocabulary. *Applied Linguistics*, *1*(24), 1–24. https://doi.org/10.1017/S0142716419000365

IBM Corp. (2019). IBM SPSS Statistics for Windows, Version 26.0. Armonk, NY: IBM Corp.

Iglesias, A., & Rojas, R. (2012). Bilingual language development of English language learners: Modeling the growth of two languages. In *Bilingual Language Development and Disorders in Spanish-English Speakers* (pp. 3–30).

Jenkins, J. R., Johnson, E., & Hileman, J. (2004). When is reading also writing: Sources of individual differences on the new reading performance assessments. *Scientific Studies of Reading*, *8*(2), 125–151. https://doi.org/10.1207/s1532799xssr0802_2

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.

Johansson, V. (2009). Lexical diversity and lexical density in speech and writing: a developmental perspective. *Working Papers in Linguistics*, *53*(0), 61–79. https://doi.org/10.7820/vli.v01.1.koizumi

Joy, R. (2011). The concurrent development of spelling skills in two languages. International Electronic Journal of Elementary Education, 3(2), 105–121.

Juel, C., Biancarosa, G., Coker, D., & Deffes, R. (2003). Walking with Rosie: A cautionary tale of early reading instruction. *Educational Leadership*, Vol. 60, pp. 12–18.

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.

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

Kettler, R. J., Glover, T. A., Albers, C. A., & Feeney-Kettler, K. A. (2014). *Universal screening in educational settings: Evidence-based decision making for schools*. American Psychological Association.

Kieffer, M. J. (2012). Early oral language and later reading development in Spanish-speaking English language learners: Evidence from a nine-year longitudinal study. *Journal of Applied Developmental Psychology*, *33*(3), 146–157. https://doi.org/10.1016/j.appdev.2012.02.003

Kim, Y. S. G. (2016). Direct and mediated effects of language and cognitive skills on comprehension of oral narrative texts (listening comprehension) for children. *Journal of Experimental Child Psychology*, *141*, 101–120. https://doi.org/10.1016/j.jecp.2015.08.003

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

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155-163.

Koutsoftas, A. D. (2013). School–age language development: Application of the five domains of language across four modalities. In N. Capone-Singleton and BB Shulman (Eds.), *Language Development: Foundations, Processes, and Clinical Applications*, 215–229. Jones & Bartlett Publishers.

Language and Reading Research Consortium. (2015). Learning to read: Should we keep things simple? *Reading Research Quarterly*, 50, 151–169. doi:10.1002/rrq.99.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*(3), 307-322.

Lee, S. H., & Muncie, J. (2006). From receptive to productive: Improving ESL learners' use of vocabulary in a postreading composition task. *TESOL Quarterly*, *40*(2), 295. https://doi.org/10.2307/40264524

Lenhard, W. & Lenhard, A. (2014). Hypothesis tests for comparing correlations. Bibergau (Germany): Psychometrica. Retrieved from https://www.psychometrica.de/correlation.html.

Lesaux, N. K., Crosson, A. C., Kieffer, M. J., & Pierce, M. (2010). Uneven profiles: Language minority learners' word reading, vocabulary, and reading comprehension skills. *Journal of Applied Developmental Psychology*, *31*(6), 475–483. https://doi.org/10.1016/j.appdev.2010.09.004

Lesaux, N. K., & Kieffer, M. J. (2010). Exploring sources of reading comprehension difficulties among language minority learners and their classmates in early adolescence. *American Educational Research Journal, 47(3),* 596-632.

Lonigan, C. J., & Milburn, T. F. (2017). Identifying the dimensionality of oral language skills of children with typical development in preschool through fifth grade. *Journal of Speech, Language, and Hearing Research*, *60*(8), 2185–2198. https://doi.org/10.1044/2017_JSLHR-L-15-0402

Lugo-Neris, M. J., Jackson, C. W., & Goldstein, H. (2010). Facilitating vocabulary acquisition of young English language learners. *Language, Speech, and Hearing Services in Schools*, *41*(3), 314–327. https://doi.org/10.1044/0161-1461(2009/07-0082)

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

Mancilla-Martinez, J., & Lesaux, N. K. (2011). The gap between Spanish speakers' word reading and word knowledge: A longitudinal study. *Child Development*, *82*(5), 1544–1560. https://doi.org/10.1111/j.1467-8624.2011.01633.x

Manis, F. R., Lindsey, K. A., & Bailey, C. E. (2004). Development of reading in grades k-2 in Spanish-speaking English-language learners. *Learning Disabilities Research and Practice*, *19*(4), 214–224. https://doi.org/10.1111/j.1540-5826.2004.00107.x

Manyak, P.C. (2012). Powerful vocabulary instruction for English learners. In E. J. Kame'enui & J. F. Baumann (Eds.), *Vocabulary instruction: Research to practice*, *second edition*. 34-50.

McKeown, M. G., Beck, I. L., & Sandora, C. (2012) Direct and rich vocabulary instruction needs to start early. In E. J. Kame'enui & J.F. Baumann (Eds.), *Vocabulary instruction: Research to practice*, 231-255. Guilford Press.

McMaster, K. L., & Campbell, H. (2008). New and existing curriculum-based writing measures: Technical features within and across grades. *School Psychology Review*, *37*(4), 550-566.

McMaster, K., & Espin, C. (2007). Technical features of curriculum-based measurement in writing: A literature review. *The Journal of Special Education*, *41*(2), 68-84.

Miller, J. F., Heilmann, J., Nockerts, A., Iglesias, A., Fabiano, L., & Francis, D. J. (2006). oral language and reading in bilingual children. *Learning Disabilities Research and Practice*, *21*(1), 30–43.

Moats, L. C. (2000). *Speech to print: Language essentials for teachers*. Paul H. Brookes Publishing Co.

Morris, N. T., & Crump, W. D. (1982). Syntactic development in the oral language of learning disabled and normal students at the intermediate and secondary level. *Learning Disability Quarterly*, *5*(2), 163–172.

Muñoz, M. L., Gillam, R. B., Peña, E. D., & Gulley-Faehnle, A. (2003). Measures of language development in fictional narratives of Latino children. *Language, Speech, and Hearing Services in Schools*, 34(4), 332–342. https://doi.org/10.1044/0161-1461(2003/027)

Nagy, W. E., & Herman, P. A. (1987). Breadth and depth of vocabulary knowledge: Implications for acquisition and instruction. *The Nature of Vocabulary Acquisition*, *19*, 35.

Nagy, W., Townsend, D., Lesaux, N., & Schmitt, N. (2012). Words as tools: Learning academic vocabulary as language acquisition. *Reading Research Quarterly*, *47*(1), 91–108. https://doi.org/10.1002/RRQ.011

Nation, I. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, *63*(1), 59-82.

National Center on Intensive Intervention. (n.d.). Academic screening tools chart. Retrieved December 07, 2019, from https://charts.intensiveintervention.org/chart/academic-screening

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

New York State Education Department (2016). ELL demographics and performance. http://www.nysed.gov/common/nysed/files/programs/bilingual-ed/factfigures_2016.pdf

Nuñez, A. M., Rios-Aguilar, C., Kanno, Y., & Flores, S. M. (2016). English learners and their transition to postsecondary education. In *Higher education: Handbook of theory and research* (pp. 41-90). Springer, Cham.

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

Olinghouse, N. G., & Wilson, J. (2013). The relationship between vocabulary and writing quality in three genres. *Reading and Writing*, *26*(1), 45–65. https://doi.org/10.1007/s11145-012-9392-5

Oregon Department of Education. (2019). *The 2017-18 Oregon English language learner report*. Retrieved from https://www.oregon.gov/ode/reports-and-data/LegReports/Pages/default.aspx

Oregon Department of Education Office of Teaching, Learning, and Assessment (2019). *ELPA test specifications*. Retrieved from https://www.oregon.gov/ode/educator-resources/assessment/Documents/elpa21testspecsg2-3.pdf

Parker, R., Tindal, G., & Hasbrouck, J. (1991). Countable indices of writing quality: Their suitability for screening-eligibility decisions. *Exceptionality: A Special Education Journal*, *2*(1), 1-17.

Pearson, P. D., Hiebert, E. H., & Kamil, M. L. (2012). Vocabulary assessment. In E. J. Kame'enui & J.F. Baumann (Eds.), *Vocabulary instruction: Research to practice*, 231-255. Guilford Press.

Perfetti, C. A., & Adlof, S. (2012). One reading comprehension: A conceptual framework from word meaning to text meaning. In J. P. Sabatini, E. R. Albro, & T. O'Reilly (Eds.), *Measuring up: Advances in how to assess reading ability*. Lanham, MD: Rowman & Littlefield Education.

Petscher, Y., Kim, Y.-S., & Foorman, B. R. (2011). The importance of predictive power in early screening assessments: Implications for placement in the response to intervention framework. *Assessment for Effective Intervention*, *36*(3), 158–166.

Pullen, P. C., Tuckwiller, E. D., Konold, T. R., Maynard, K. L., & Coyne, M. D. (2010). A tiered intervention model for early vocabulary instruction: The effects of tiered instruction for young students at risk for reading disability. *Learning Disabilities Research & Practice*, *25*(3), 110–123. https://doi.org/10.1111/j.1540-5826.2010.00309.x.

R Development Core Team, R. (2011). *R: A language and environment for statistical computing*. *R Foundation for Statistical Computing*. https://doi.org/10.1007/978-3-540-74686-7

Read, J. (2000). *Assessing vocabulary* (pp. 1-85). Cambridge: Cambridge university press.

Read, J., & Chapelle, C. A. (2001). A framework for second language vocabulary assessment. *Language Testing*, *18*(1), 1-32.

Ritchey, K. D., & Coker Jr, D. L. (2013). An investigation of the validity and utility of two curriculum-based measurement writing tasks. *Reading & Writing Quarterly*, *29*(1), 89-119.

Romig, J. E., Therrien, W. J., & Lloyd, J. W. (2017). Meta-Analysis of Criterion Validity for Curriculum-Based Measurement in Written Language. *Journal of Special Education*, *51*(2), 72–82. https://doi.org/10.1177/0022466916670637.

Salvia, J., Ysseldyke, J. E., & Witmer, S. (2017). Assessment in Special and Inclusive Education (13th ed.). Boston, MA: Cengage.

Scott, C. (2009). Language-Based Assessment of Written Expression. In Troia, G. A. (Ed.), *Instruction and assessment for struggling writers: Evidence-based practices*. Guilford Press.

Shanahan, T. (2006). Relations among oral language, reading, and writing development. In C. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research*, 171-183.

Silverman, R. D., Coker, D., Proctor, C. P., Harring, J., Piantedosi, K. W., & Hartranft, A. M. (2015). The relationship between language skills and writing outcomes for linguistically diverse students in upper elementary school. *The Elementary School Journal*, *116*(1), 103–125.

Silverman, S., & Ratner, N. B. (2002). Measuring lexical diversity in children who stutter: Application of vocd. *Journal of Fluency Disorders*, *27*(4), 289–304. https://doi.org/10.1016/S0094-730X(02)00162-6

Simon-Cereijido, G., & Gutierrez-Clellen, V. F. (2009). A cross-linguistic and bilingual evaluation of the interdependence between lexical and grammatical domains. *Applied Psycholinguistics*, *30*(2), 315–337. https://doi.org/10.1017/S0142716409090134

Smith, R. A., & Lembke, E. S. (2020). Aspects of technical adequacy of an early-writing measure for English language learners in grades 1 to 3. *Assessment for Effective Intervention*, *1*(5). https://doi.org/10.1177/1534508420947157

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, *36*(2), 139–152. https://doi.org/10.1080/09571730802389975

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

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, 1, 1–26.

Thordardottir, E., Rothenberg, A., Rivard, M.-E., & Naves, R. (2006). Bilingual assessment: Can overall proficiency be estimated from separate measurement of two languages? *Journal of Multilingual Communication Disorders*, *4*(1), 1–21. https://doi.org/10.1080/14769670500215647

Tindal, G. (2013). Curriculum-based measurement: A brief history of nearly everything from the 1970s to the present. *International Scholarly Research Notices*, *2013*.

Tindal, G., & Parker, R. (1989). Assessment of written expression for students in compensatory and special education programs. *The Journal of Special Education*, 23, 169–183.

Troia, G. A. (2009). Introduction. In G. A. Troia (Ed.), *Instruction and assessment for struggling writers: Evidence-based practices*. Guilford Press.

Uccelli, Paola; Páez, M. (2007). Narrative and vocabulary development of bilingual children from kindergarten to first grade: developmental changes and associations among English and Spanish skills. *Language, Speech, and Hearing Services in Schools*, *38*, 225–236.

Uchikoshi, Y., Yang, L., Lohr, B., & Leung, G. (2016). Role of oral proficiency on reading comprehension. *Literacy Research: Theory, Method, and Practice*, *65*(1), 236–252. https://doi.org/10.1177/2381336916661538

Umansky, I. M. (2016). To be or not to be EL: an examination of the impact of classifying students as English learners. *Educational Evaluation and Policy Analysis*, *38*(4), 714–737. https://doi.org/10.3102/0162373716664802

University of Oregon Center on Teaching and Learning (2019). *Instructional Grouping Worksheets, DIBELS 8th Edition*. Retrieved from: https://dibels.uoregon.edu/docs/marketplace/dibels/worksheets/DIBELS-8th-IGW-Instructions.pdf

U.S. Department of Education Office for Civil Rights (2015). Dear Colleague Letter, English Learner Students and Limited English Proficient Parents (01/7/2015). https://www2.ed.gov/about/offices/list/ocr/letters/colleague-el-201501.pdf

VanDerHeyden, A. M., & Burns, M. K. (2018). Improving Decision Making in School Psychology: Making a Difference in the Lives of Students, Not Just a Prediction About Their Lives. *School Psychology Review*, *47*(4), 385–395. https://doi.org/10.17105/spr-2018-0042.v47-4

Watkins, R. V., Kelly, D. J., Harbers, H. M., & Hollis, W. (1995). Measuring children's lexical diversity: Differentiating typical and impaired language learners. *Journal of Speech and Hearing Research*, *38*(6), 1349–1355. https://doi.org/10.1044/jshr.3806.1349

Wise, J. C., Sevcik, R. A., Morris, R. D., Lovett, M. W., & Wolf, M. (2007). The relationship among receptive and expressive vocabulary, listening comprehension, pre-reading skills, word identification skills, and reading comprehension by children with reading disabilities. *Journal of Speech, Language, and Hearing Research*, *50*(4), 1093-1109.

Wood, C., Wofford, M. C., Gabas, C., & Petscher, Y. (2018). English narrative language growth across the school year: Young Spanish–English dual language learners. *Communication Disorders Quarterly*, *40*(1), 28–39. https://doi.org/10.1177/1525740118763063

Woolpert, D. (2016). Doing more with less: the impact of lexicon on dual-language learners' writing. *Reading and Writing*, *29*(9), 1865-1887.

WIDA Consortium. (2012). 2012 Amplification of the English language development standards kindergarten-grade 12. Board of Regents of the University of Wisconsin System, Madison, WI.

Youngstrom, E. A. (2014). A primer on receiver operating characteristic analysis and diagnostic efficiency statistics for pediatric psychology: We are ready to ROC. *Journal of Pediatric Psychology*, *39*(2), 204-221.

Yu, G. (2010). Lexical diversity in writing and speaking task performances. *Applied Linguistics*, *31*(2), 236–259. https://doi.org/10.1093/applin/amp024

Zareva, A., Schwanenflugel, P., & Nikolova, Y. (2005). Relationship between lexical competence and language proficiency: Variable sensitivity. *Studies in Second Language Acquisition*, *27*(4), 567–595. https://doi.org/10.1017/S0272263105050254