Article

Hammill Institute on Disabilities

Learning Disability Quarterly
2022, Vol. 45(2) 144–155

DOI: 10.1177/0731948720930423
journals.sagepub.com/home/ldq
SAGE

Technical Adequacy of a Spelling Curriculum-Based Measure for English Language Learners in the First Through Third Grades

R. Alex Smith, PhD[1], and Erica S. Lembke, PhD[2]

Abstract

This study represents an initial examination of the potential of word dictation (WD), a form of curriculum-based measure, to identify risk in writing for young English language learners (ELs). The sample included 73 ELs with beginning to advanced English language proficiency in the first to third grades attending schools in one U.S. Midwestern school district. This district used an English-only model of instruction and students completed two forms of WD in the fall, winter, and spring. The criterion measure, a common English language proficiency assessment, was administered in the winter. Results indicated that WD maintained similar reliability and validity coefficients between the ELs in this study and previous studies with the general population. Implications for practice and future research are discussed.

Keywords

writing assessment, English learners, early elementary screening

English language learners (ELs) represent a growing and extremely diverse subpopulation of students within U.S. schools (Kena et al., 2016). Currently, ELs represent approximately 17% of the first graders (Aud et al., 2012; Kena et al., 2016) and are the fastest-growing subpopulation of students (Aud et al., 2012). Although often presented as one homogeneous subpopulation, ELs are extremely diverse in terms of English language proficiency (ELP), culture, economic status, educational background, and they speak more than 400 native languages (Abedi & Gándara, 2006; National Academies of Sciences, Engineering, and Medicine [NASEM], 2017). In addition to representing a growing and diverse group of students, ELs are also at increased risk of dropping out of school and not attaining a high-school diploma (Fry, 2010). Furthermore, recent data from the National Assessment of Educational Progress (NAEP) indicate that ELs have consistently performed below their native English-speaking (NES) peers across content areas (Kena et al., 2016). Most notably, the discrepancy between ELs and their NES peers in writing performance exceeds that of either reading or mathematics (National Center for Education Statistics [NCES], 2012).

Writing is a critical aspect of long-term literacy development. Writing ability affects a student's success across content areas in K–12 as well as access to and success in both postsecondary education and vocational settings (Abbott et al., 2010; Berninger & Abbott, 2010; Graham & Hebert, 2011; Graham & Perin, 2007; Mueller & Oppenheimer, 2014). As noted previously, while the academic performance of ELs is generally behind that of their NES peers, this discrepancy is even wider for tasks with higher levels of language demand like writing (Abedi & Gándara, 2006). NAEP reported ELs in 12th grade performed lower in writing than any other subpopulation of students, including those with disabilities (NCES, 2012). Fortunately, early identification of risk via reliable and valid assessments coupled with access to evidence-based early interventions has been shown to improve student outcomes in writing (Berninger & Amtmann, 2003; Graham et al., 2001). However, there is little research related to assessment for early identification of risk in writing specifically for ELs (Keller-Margulis et al., 2016; Miller & McCardle, 2011). The lack of research and available assessment tools has translated into a delay in receiving access to intervention resources during critical early stages of literacy development for many young ELs (NASEM, 2017).

[1]The University of Southern Mississippi, Hattiesburg, USA
[2]University of Missouri, Columbia, USA

Corresponding Author:
R. Alex Smith, School of Education, The University of Southern Mississippi, 118 College Drive, Hattiesburg, MS 39406, USA.
Email: r.alexander.smith@usm.edu

Early Identification

Research has documented the rates of underidentification of ELs with learning disability (LD) in the early elementary grades followed by patterns of overidentification by the upper elementary grades (NASEM, 2017). Estimates of the prevalence of LD in writing range from 6.9% to 14.7% (Katusic et al., 2009), along with high rates of comorbid writing and reading LD (Costa et al., 2016). Considering that ELs may be at increased risk in writing as compared with other domains (Abedi & Gándara, 2006), more research informing early identification and access to appropriate resources (e.g., supplemental intervention or special education) is critical. Currently, a response to intervention (RTI) process has been identified as the most reliable and valid framework for identifying LD (Fletcher & Miciak, 2019). Researchers have also underscored the promise of RTI in promoting a more culturally responsive method for identifying ELs with LD (Burr et al., 2015; O'Connor et al., 2014). Furthermore, RTI has shown promise not only for promoting early identification of LD but also for promoting the academic growth of all students via early identification of risk and early intervention (Burr et al., 2015; O'Connor et al., 2014). RTI entails early identification of risk via screening and then measuring the at-risk individual's responsiveness (i.e., subsequent academic growth) to evidence-based practices via progress monitoring (Fletcher & Miciak, 2019). However, valid and reliable assessments for screening and progress monitoring are necessary features of RTI and subsequent identification of LD (Burr et al., 2015).

One form of assessment commonly used within RTI for both universal screening and progress monitoring is curriculum-based measurement (CBM; Deno, 1985). CBMs are considered general outcome measures that assess developmentally appropriate skills aligned with grade-level curriculum that are reliable, valid, sensitive to growth, and are relatively easy to administer and score (Deno, 1985). Fuchs (2004) outlined three stages of research needed in "substantiating the tenability" (p. 189) of CBM within RTI. Stage 1 is the technical adequacy of the static score, Stage 2 examines the technical features of slope/growth, and Stage 3 examines the instructional utility of CBM, that is, whether teachers can use the data in meaningful ways. Each stage of research expounded by Fuchs has different priorities in relation to the multidimensional construct of validity.

Validity is the extent to which assessment data and theory converge to support a test's hypothesized purpose and its application in practice (Salvia et al., 2017). Stage 1 examines validity of the static score for the purpose of screening. Validity is constrained by reliability (i.e., a valid assessment must be reliable whereas an unreliable assessment is not valid; Salvia et al., 2017). *Reliability* refers to the extent to which similar scores are generated across items, time, and people (Salvia et al., 2017). The standards for reliability provided by Salvia and colleagues (2017) include $r \geq .70$ for progress monitoring, $r \geq .80$ for screening, and $r \geq .90$ for high-stake decisions (i.e., special education placement). However, validity goes beyond reliability alone.

Validity refers to whether the assessment data result in correct inferences "about a specific person in a specific situation for a specific purpose" (p. 79, Salvia et al., 2017). In regard to Stage 1, the purpose of screening is risk identification and the need for additional data for risk confirmation (Fuchs, 2004). Risk should be determined in accordance with socially valid outcomes (i.e., outcomes that have meaningful implications to the individual), whereas the screener's ability to predict risk should be examined via both theoretical and quantitative evidence (Messick, 1989). Common methods of examining the validity of a screener include convergent validity (i.e., the extent to which the screener correlates to a criterion measure of the same or related construct), divergent validity (i.e., the extent to which scores on a screener do not correlate to a measure of an unrelated construct), and content validity (i.e., the extent to which screener items represent the construct being assessed; Messick, 1989). Validity is an on-going and iterative process of examining assessment to provide practitioners a wealth of timely and relevant data by which they may make informed decisions related to the intended purpose of assessment and target population (Salvia et al., 2017). Furthermore, although an assessment may provide data that are deemed highly reliable and valid, it is still best practice to use multiple data sources for risk certification and especially for LD identification (Fletcher & Miciak, 2019).

As validity pertains specifically to ELs, Abedi and Gándara (2006) recommend that practitioners select assessments that have been created and normed specifically for/with ELs to reduce bias. In line with Abedi and Gándara's recommendation, it is best practice to initially examine validity of an assessment using a criterion measure that not only has social relevance specifically for ELs but also has been created specifically for and normed with ELs. All states require ELP assessments for ELs receiving English as second or other language services (ESOL; Fox & Fairbairn, 2011). Furthermore, performance on these ELP assessments determines whether or not an EL receives ESOL services, often guides the nature of ESOL services, and is used for accountability purposes (Fox & Fairbairn, 2011). Therefore, ELP assessments are socially valid criterion measures that were created specifically for ELs. However, there are several other variables specific to ELs that may influence the validity of assessment and how assessment data should be interpreted.

ELs receive instruction via an English-only or a bilingual model (i.e., instruction delivered in both English and the student's native language), although the nature of bilingual models can vary substantially (Genesee et al., 2005; NASEM, 2017). Research indicates that ELs receiving bilingual instruction exhibit initial lags in growth in English literacy as

compared with their peers receiving English-only instruction but later demonstrate accelerated rates of growth and eventually outperform their peers in English-only models (Genesee et al., 2005; NASEM, 2017). In short, the model of instruction influences expected rates of growth as well as the long-term predictive validity of a given static score. Another key variable in establishing the technical adequacy of assessment for ELs is oral English proficiency (Keller-Margulis et al., 2012; NASEM, 2017). For example, Keller-Margulis and colleagues (2012) found that the validity of static scores and patterns of seasonal growth on CBM-Reading varied according to an EL's initial ELP. An EL's ELP is generally classified as beginning, intermediate, or advanced according to an ELP assessment. An EL in the early stages of English language acquisition (i.e., beginning ELP) should not be expected to perform at the same level or learn English literacy skills at the same rate as an NES peer or an EL peer with more advanced ELP. Unfortunately, the majority of standardized assessments only report performance data for ELs as a single homogeneous subpopulation and do not provide data specific to model of instruction or ELP (Abedi & Gándara, 2006).

Early Identification of Risk and Assessment in Writing

Written expression has proven difficult to measure (Salvia et al., 2017). For example, the majority of standardized assessments in writing fail to exceed $r \geq .60$ for criterion-related validity and have a wide range of reliability (Taylor, 2006). Assessment of writing typically falls below the domains of content or form (Salvia et al., 2017). Content assessments often use rubrics and focus upon word choice, organization, clarity, and/or complexity. Unfortunately, content assessment is prone to concerns with reliability (McMaster, Ritchey, et al., 2011; Salvia et al., 2017). Assessments of form relate to the mechanics or rules of writing (e.g., letter formation, spelling, grammar), are easier to objectively measure than content-based assessment, and constitute the majority of diagnostic assessments in writing (Salvia et al., 2017). How form is assessed (i.e., handwriting, spelling, or grammar) changes as a function of the individual's stage in writing development (Salvia et al., 2017). Age- and grade-level curricular expectations are generally used as a proxy for an individual's expected stage in writing development, though ELP should also be considered in regard to an EL's English writing. For example, a high-school-age EL with beginning ELP may still be working on entry-level transcription skills (e.g., basic spelling) despite grade-level expectations.

Writing consists of multiple processes and demands the coordination of several skills (e.g., speaking, listening, reading) to produce quality composition (McMaster, Du, et al., 2011; Salvia et al., 2017). Writers develop these processes and build the various skills across their lifespan, but transcription and ideation are the most influential for beginning writers (McMaster, Ritchey, et al., 2011). Theoretical models of early writing, such as the "Simple View" (Berninger et al., 2002) and later the "Not So Simple View" (Berninger & Winn, 2006), explicate how building fluency in transcription skills (i.e., handwriting, spelling) allows finite cognitive resources (e.g., working memory) to be reallocated to support text generation (i.e., thinking of the words and phrases to translate into text) and executive functioning (e.g., monitoring attention, planning, revising; Berninger et al., 2002; Berninger & Swanson, 1994; Berninger & Winn, 2006).

Transcription skills explain a large amount of variance in both the quantity and the quality of composition of writers from as early as first grade to as late as ninth grade (Abbott et al., 2010; Graham et al., 1997). According to Abbott and colleagues (2010), "spelling explained unique variance in both word-level spelling and text-level composition consistently from first through seventh grades" (p. 294). Furthermore, research has demonstrated that interventions directly targeting transcription skills for beginning writers have promoted improvement in students' overall writing quality (Berninger et al., 1997; Graham et al., 2002; Jones & Christensen, 1999), which suggests a causal relationship (McMaster, Ritchey, et al., 2011). In addition, research indicates that students with writing LD often struggle specifically with transcription skills (Graham et al., 2001).

CBM-Writing

CBM-Writing has been researched with the general population across Stage 1 (McMaster & Espin, 2007; Romig et al., 2017), Stage 2 (McMaster, Du, et al., 2011), and Stage 3 (McMaster et al., 2020). Due to issues inherent with assessment in writing, such as variable reliability and low criterion validity between standardized assessments (Salvia et al., 2017; Taylor, 2006), correlations of $r \geq .50$ with a criterion measure are generally accepted as adequate in CBM-Writing research for screening and progress monitoring (McMaster, Ritchey, et al., 2011). In general, CBM-Writing has been found to be a reliable and valid predictor of future writing performance (Romig et al., 2017), but there are several types of CBM-Writing, and the technical adequacy of each type varies in accordance with the student's grade-level curricular expectations and stage of writing development (McMaster, Ritchey, et al., 2011). CBM-Writing researchers have leveraged the theoretical and intervention research with beginning writers to develop appropriate forms of early CBM-Writing (McMaster, Ritchey, et al., 2011). Although the most common form of CBM-Writing is story prompt (i.e., a passage-level measure on which a student composes a story within a set amount of time), it has produced insufficiently consistent reliability ($r = .20–.47$) and criterion validity ($r = .23–.63$) for beginning writers (McMaster, Ritchey, et al., 2011). One form of CBM-Writing that has shown promise for beginning writers is word dictation (WD).

WD assesses a student's word-level transcription skills (i.e., letter formation and spelling) by asking the student to spell as many dictated words as they can within a set amount of time. The student's written responses on WD are then scored using a variety of metrics, including total number of words written (WW), total number of words spelled correctly (WSC), total correct letter sequences (CLS), and correct minus incorrect letter sequences (C-ILS). CLS are defined as any two adjacent correct letters in accordance with the dictated word (McMaster & Lembke 2014). Metrics are often categorized as production-only (i.e., WW) and accurate-production (e.g., WSC, CLS). Percentage metrics (i.e., %WSC) have also been examined and are generally classified as production independent. In comparing several different types of CBM-Writing, Lembke and colleagues (2003) found that dictation CBM-Writing, including WD, had the strongest criterion validity in the second grade when compared with copying CBM-Writing.

WD has demonstrated adequate reliability (r = .89–.95) and validity (r = .29–.75), and it is adequate for weekly or biweekly progress monitoring for students in the first to third grades (Hampton & Lembke, 2016; Lembke et al., 2003; Poch et al., 2019). Ritchey and Coker (2014) also found WD was adequate in identifying risk in the first grade (area under the curve [AUC] = .78–.87). In a study with first- through third-grade students, Poch and colleagues (2019) examined predictive (fall to spring) and concurrent (spring to spring) validity of WD. Validity coefficients for first grade ranged from r = .11 to .53, but accurate production metrics ranged from .37 to .53. Validity for second and third grades ranged from r = .48 to .77, with accurate-production metrics (r = .59–.77) again producing higher coefficients than WW. A study by Keller-Margulis and colleagues (2019), though not specifically with ELs, examined a form of WD with kindergartners who were approximately 50% EL. They found concurrent validity coefficients with a standardized assessment in writing ranging from .26 to .70, whereas WSC (r = .60–.70) and CLS (r = .56–.68) had consistently higher coefficients than WW. However, to our knowledge, no studies have been published in an academic journal that examine WD specifically with ELs in any grade level, and only one study (Keller-Margulis et al., 2016) has examined any form of CBM-Writing exclusively with elementary-age ELs (fourth grade).

Keller-Margulis and colleagues (2016) examined story prompt with fourth-grade students (N = 139), of which 19 were ELs, and they conducted separate analyses for the ELs in their study. Although ELP and native language were not reported, ELs received a bilingual model of ESOL instruction (approximately 75% of instruction in English by fourth grade; Keller-Margulis et al., 2016). The criterion measure used was the state of Texas writing assessment. Correlations were found with various story prompt metrics, and the criterion measure ranged from r = −.04 to .68, with accurate-production and production-independent metrics producing the most significant correlations. Keller-Margulis and colleagues concluded that story prompt may not be adequate for screening young ELs, at least for the sample included in their study.

However, Keller-Margulis and colleagues' (2016) findings are in contradiction to findings by other studies examining story prompt with older ELs with intermediate or better ELP, which consistently found validity coefficients equitable to or better than those found for NES peers (Campbell et al., 2013; Espin et al., 2008). The only study including ELs with beginning ELP included high-school-age ELs and found that a passage copying CBM-Writing produced reliability and validity coefficients that were comparable to those found with younger (early elementary) NES peers (Campbell, 2010). Therefore, the existing literature suggests that the most appropriate form of CBM-Writing for ELs changes as a function of both grade/age and ELP. Passage-level CBM-Writing (i.e., story prompt) may be inappropriate for young ELs (fourth grade or below; Keller-Margulis et al., 2016) and for ELs with beginning ELP regardless of age (Campbell, 2010). However, Lembke and colleagues (2003) found that dictation forms of CBM-Writing had higher validity coefficients than copying forms of CBM-Writing for young writers in general, which indicates that WD may be more appropriate for young ELs as well as for ELs with beginning ELP. Furthermore, research by Keller-Margulis and colleagues (2019) indicates the potential of WD to be adequate for screening purposes with young ELs.

Purpose of the Present Study

WD may be a more appropriate form of CBM-Writing for young ELs because it is directly aligned with early writing instruction and curricula (i.e., word-level transcription skills) while also maintaining long-term predictiveness of overall writing quality for all students (Abbott et al., 2010) as well as specifically for ELs (Harrison et al., 2016). Furthermore, spelling deficits have been found to be highly predictive of both reading and writing LD within the general population (Berninger et al., 2008). Therefore, the purpose of this study is to provide an initial examination (i.e., Stage 1) of WD for ELs in the first through third grades. The specific research questions guiding the study were:

Research Question 1: How does the reliability of WD for young ELs compare to the standards set by Salvia and colleagues (2017)?

Research Question 2: How does the concurrent and predictive validity of WD for young ELs compare to prior research on WD and CBM-Writing with the general population?

Table 1. Student Participant Demographics.

Grade	*n*	Male (%)	#ELP Beg/Int/Ad	Spanish (%)	Arabic (%)
1	24	67	6/15/2[a,b]	41.7	20.8
2	25	60	7/18/0	28	4
3	24	67	8/15/0[a]	45.8	4.2
Total *N*	73	64	21/48/2[a,b]	38.4	9.6

Note. %Spanish and %Arabic provide the percentage of sample speaking each respective language as their native language. ELP = English language proficiency; Beg = beginning proficiency; Int = intermediate proficiency; Ad = advanced proficiency.
[a]One student did not have available ELP data to report in Grades 1 and 3.
[b]One advanced proficiency ELP participant removed from validity analysis as an outlier.

Method

Recruitment and Consent

One U.S. Midwestern school district serving a mid-sized city agreed to participate in the study. The researchers first sought consent from district ESOL teachers ($N = 17$), of which 10 consented to participate and remained throughout the study, for a 59% participation rate. Of the 15 elementary school buildings with ESOL teachers, participating ESOL teachers were in 9 for a 60% school participation rate. The estimated total number of EL students in Grades 1–3 across the district was 380, of which parental consent letters were sent home to approximately 230, with letters to the EL students of participating ESOL teachers only. Seventy-three students returned parental consent, returned signed student assent to participate in the study, and completed the study for a participation rate of 32% of those receiving parental consent letters and approximately 19% across all first- to third-grade ELs in the district. Consent letters were provided to ELs in both English and their native language by the researchers in the presence of their ESOL teacher.

Setting and Participants

All participants received English-only ESOL instruction. The district served 18,000 K–12 students during the 2016–2017 academic year. Across the district, students were 60.8% White, 20% Black, 6.3% Hispanic, and 5.5% Asian. Furthermore, 39.7% of students were eligible for free and reduced-price lunch, 9.7% had IEPs, and 6% received ESOL services during the 2016–2017 academic year. The demographics of the specific schools that participants attended were 52.1% free and reduced-price lunch, 9.8% with IEPs, and 12% receiving ESOL services.

Student participants. The total number of student participants in this study was 73 ELs across Grades 1–3. Table 1 provides sample size by grade and information regarding gender, ELP, and native language. Participants were considered beginning if their overall ELP proficiency score was 2.9 or less, intermediate if their score was 3.0–4.9, and advanced if the student scored above 4.9. Beyond the two most common native languages (i.e., Spanish and Arabic), other languages represented were Burmese, Tigrinya, Korean, Vietnamese, Czech, Karenni, Chinese, Kirundi, Tagalog, and Swahili. Two students did not take or had incomplete data on the 2016–2017 ELP assessment. Thus, only 71 of 73 participants had completed ELP data to report and were included in validity analyses.

Data collectors. Six data collectors assisted with the distribution and collection of assent/consent forms as well as with administration and scoring of WD. Five of the six data collectors had 3 or more years of experience administering and scoring WD. One data collector was an undergraduate research assistant who was trained in WD assessment until he achieved 90% fidelity of administration. All data collectors were native English speakers.

Measures

Word dictation. The WD forms used for this study were created by McMaster and colleagues (2014). These WD forms were selected because they had published grade-level norms (see McMaster & Lembke 2014) and have been previously examined in published peer-reviewed research with the general population (see Poch et al., 2019). According to McMaster and colleagues, the researchers generated 30 forms of WD, the words were selected based on the Common Core State Standards in the early elementary grades, and each form featured words progressing from less complex (e.g., CVC) to more complex (e.g., CCVVC). WD was administered and scored according to the standard procedures published by McMaster et al. WD is individually administered for 3 min, the administrator dictates each word twice, and the student is to write as many of the 30 words as possible in 3 min. In the case that a student completes all 30 words in less than 3 min, scores are prorated by calculating a score per second and then multiplying by 180 s. The specific metrics examined included WW, WSC, CLS, ILS, C-ILS, %WSC, and %CLS.

Criterion measure. The Assessing Comprehension and Communication in English State-to-State ELP (ACCESS) test is the most common ELP assessment across states (Fox & Fairbairn, 2011). The ACCESS test was used as a criterion measure because it was specifically developed for and normed using ELs (Fox & Fairbairn, 2011). The ACCESS consists of four subtests: Reading, Writing, Speaking, and Listening. The ACCESS Writing subtest (ACCESS-W) was used as the primary criterion measure for all ELs in this study. The Reading and Writing subtests of ACCESS are the strongest predictors of performance on content area state

tests (Parker et al., 2009). In addition, the ACCESS-W significantly predicted ($p<.001$) performance on the writing subtest of the New England Common Assessment Program for both fifth- and eighth-grade ELs (Yanosky et al., 2011).

Time allocated for completion of the ACCESS-W is 35 min for first graders and 65 min for third graders and is scored by trained professionals outside of the district using established rubrics, underscoring the purpose of ACCESS as a diagnostic/summative assessment not meant for screening or progress monitoring. For the ACCESS-W, students are read a variety of prompts/scripts with accompanying images to which they then compose a variety of responses. Spelling, capitalization, and punctuation are not counted negatively, and scoring is content based. The scoring rubric used assesses writing according to discourse (i.e., organization and appropriateness to context), sentences (i.e., variety, complexity, grammar), and words (i.e., range of vocabulary and appropriateness to context). (More information regarding the additional subtests of the ACCESS are available via e-mail to the lead author).

Procedures

Participants took two alternate forms (A and B) of WD on the same day in the fall (mid-November), winter (late-January), and spring (mid-April). The forms were counterbalanced across students and time points. For validity, the mean of scoring metrics was used. WD was administered within a 2-week window across teachers and schools for each benchmark. All participants completed the winter WD benchmark within 2 weeks of completing the ACCESS-W. Each benchmark was separated by a minimum of 8 weeks, with an average of 9 weeks between fall and winter and an average of 11 weeks between winter and spring. Approximately 10% of all WD administrations per administrator were evaluated for fidelity of administration by a trained observer using a modified version of the Accuracy of Implementation Rating Scales (Fuchs et al., 1984). Total fidelity of administration was 99%. All scored data were double counted and entered, and any discrepancies were discussed and remediated on an individual basis. This process accounted for any counting and/or data entry errors.

Data Analysis

Reliability. A random selection of 24% of WD across the three time points was double scored for inter-scorer reliability (ISR). The total percentage of WD scored in fall, winter, and spring time points for ISR was 22%, 26%, and 22%, respectively. Scoring reliability was calculated by dividing total scoring agreements by agreements plus disagreements. ISR was calculated for WW, WSC, CLS, and ILS. Pearson's correlations were used between forms A and B within each grade for each time point to examine alternate-form reliability.

Validity. Predictive and concurrent validity were explored using Pearson's correlations within each grade between WD metrics and the ACCESS-W. Correlations are also provided for the various subtests of the ACCESS to examine divergent validity for any metrics meeting minimum requirements for adequacy (i.e., reliability of $r \geq .80$ and validity of $r \geq .50$).

Results

Descriptive Statistics

Criterion measure. Skew and kurtosis were acceptable for ACCESS-W for second and third grades but was significantly skewed for first grade ($|z| > 1.96$, $p < .05$). Shapiro–Wilk also indicated a non-normal distribution ($p < .05$) for first grade. Further analysis of the extreme outlier indicated that Participant 62 was a first-grade student with an overall composite score on the ACCESS as advanced. Participant 62 had the highest score across all grades on the ACCESS-W. Therefore, that student's results were removed and descriptive statistics were run again for first grade. ACCESS-W was normally distributed according to skew and kurtosis once the outlier was removed. Shapiro–Wilk also indicated that the ACCESS-W was normal ($p = .76$) once the outlier was removed. The decision was made to remove Participant 62 from further validity analysis. More detailed descriptive data regarding the additional subtests of the ACCESS are beyond the scope of this article but are available from the lead author.

Predictor measure. Descriptive statistics are provided for WW, WSC, CLS, %WSC, %CLS, and C-ILS in Table 2. In general, a value within ± 1.96 when dividing the skew or kurtosis by its respective standard error to convert it to a z score indicates a normal distribution (Ghasemi & Zahediasl, 2012), and this method was used to identify significance in both skew and kurtosis. Box plots and histograms were also used to identify potential outliers, and as Shapiro–Wilk was used to test for normality. There were consistent concerns related to both skew and kurtosis in first grade, especially in winter and spring, for %CLS and C-ILS. %CLS was significantly skewed in the spring across grade levels. WSC was also significantly skewed in fall and winter for first grade. However, all metrics were retained for further analyses.

Reliability

ISR was 94% or better across all metrics (i.e., WW, WSC, CLS, ILS) with a range of 94–99%. Alternate-form reliability for each time point, grade, and scoring procedure is provided in Table 3. ILS was the only scoring procedure not meeting the alternate-form threshold of .70 (progress monitoring) for second and third graders in the fall and third

Table 2. Word Dictation Descriptive Statistics.

Grade and Metric	Fall				Winter				Spring			
	M (*SD*)	Skew	Kurtosis	Range	*M* (*SD*)	Skew	Kurtosis	Range	*M* (*SD*)	Skew	Kurtosis	Range
First grade												
WW	17 (7)	0.09	−0.95	5,29	20 (7)	0.49	−0.69	10,32	23 (8)	0.67	0.45	9,43
WSC	5 (5)	1.22*	1.41	0,18	6 (5)	1.24*	0.54	1,18	8 (7)	0.95	−0.26	0,22
CLS	49 (28)	0.59	−0.52	14,111	57 (29)	0.68	−0.36	15,120	67 (36)	0.61	−0.12	7,146
%WSC	0.26 (0.17)	0.39	−0.43	0,66	0.29 (0.19)	0.91	0.47	5,75	0.34 (0.23)	0.43	−1.11	0,71
%CLS	0.65 (0.12)	−0.74	0.33	34,84	0.64 (0.17)	−1.39*	2.41*	14,87	0.67 (0.19)	−2.04*	6.35*	4,88
C-ILS	25 (26)	0.67	0.49	−19,90	26 (41)	−1.04*	3.11*	−96,89	33 (56)	−1.89*	6.92*	−160,117
Second grade												
WW	23 (9)	−0.26	1.04	2,44	27 (9)	0.18	−0.47	15,56	31 (11)	0.59	−0.43	1,33
WSC	12 (10)	0.69	−0.69	1,33	15 (11)	0.37	−1.13	3,35	19 (13)	0.58	−0.61	4,44
CLS	86 (49)	0.37	−0.47	6,195	105 (52)	0.34	−0.94	31,211	123 (63)	0.64	−0.70	46,248
%WSC	0.47 (0.29)	0.27	−1.19	7,95	0.52 (0.28)	−0.04	−1.27	11,93	0.60 (0.27)	−0.36	−0.94	11,98
%CLS	0.73 (0.16)	−0.29	−0.48	39,99	0.76 (0.17)	−0.76	0.05	36,98	0.79 (0.17)	−1.28*	1.29	34,98
C-ILS	61 (55)	0.47	−0.93	−21,166	76 (61)	0.05	-0.94	−40,184	94 (73)	0.06	−0.47	−50,122
Third grade												
WW	26 (8)	0.30	−0.07	14,42	31 (7)	−0.04	−0.15	17,45	33 (9)	0.48	−0.03	18,54
WSC	15 (8)	0.27	−0.94	2,29	17 (8)	0.12	−0.93	3,34	21 (10)	0.54	−0.01	6,44
CLS	103 (40)	0.31	−0.39	42,178	121 (40)	0.07	−0.35	38,203	138 (49)	0.55	0.02	65,201
%WSC	0.53 (0.19)	−0.29	−0.83	13,86	0.54 (0.21)	−0.51	−0.95	15,80	0.61 (0.17)	−0.71	0.43	23,90
%CLS	0.77 (0.10)	−0.79	−0.05	54,93	0.77 (0.12)	−0.88	−0.01	48,92	0.82 (0.09)	−1.35*	2.10*	58,95
C-ILS	76 (42)	0.09	−0.87	6,152	86 (47)	0.01	−0.80	−4,177	110 (52)	0.38	0.09	21,228

Note. WW = words written; WSC = words spelled correctly; CLS = correct letter sequences; ILS = incorrect letter sequences; C-ILS = correct minus incorrect letter sequences.
*$p < .05$ when skew/kurtosis divided by standard error is greater than ± 1.96.

graders in the spring. Beyond ILS, the coefficients ranged from $r = .73$ to .97, indicating that all metrics besides ILS met the threshold of .70 for alternate-form reliability for progress monitoring. However, according to standards set by Salvia and colleagues (2017) for screening, only WW, WSC, CLS, and C-ILS met the $r \geq .80$ standard for each time point and grade.

Validity

Concurrent validity was examined by correlating winter WD to ACCESS-W, and predictive validity was examined via correlating fall WD metrics to ACCESS-W. Results are found in Table 4. WSC, CLS, %WSC, and C-ILS met the $r \geq .50$ criteria for all time points examined across all grades with a range of $r = .52$–.81. In general, accurate production (e.g., CLS, WSC) was more consistent across grades and time points in comparison to percentage metrics and either outperformed or performed on par with WW. Divergent validity was examined by reporting correlations between promising WD metrics, according to both reliability and convergent validity, and the various subtests of the ACCESS, including Reading, Speaking, and Listening. The metrics included were WSC, CLS, and C-ILS. Percentage WSC was not included because of concerns with alternate-form reliability. Results are available in Table 5. Correlations were higher for the ACCESS-W than any other subtest across grades and metrics except for WSC and C-ILS for first grade in the winter.

Discussion

The purpose of this study was an initial examination of the potential for WD to identify risk. Many WD metrics, particularly the accurate-production metrics, met the reliability standards for screening ($r \geq .80$) set forth by Salvia and colleagues (2017). Accurate-production WD metrics also met the $r \geq .50$ criterion validity coefficients commonly accepted within the CBM-Writing literature (McMaster, Ritchey, et al., 2011) and aligned with prior research employing WD with the general population (Keller-Margulis et al., 2019; Poch et al., 2019). Accurate-production metrics displayed consistently higher coefficients with the ACCESS-W than the other ACCESS subtests across second and third grades whereas CLS did so across all three grades. Although correlations were often significant across multiple subtests (e.g., Writing and Reading), this is to be expected because writing incorporates skills from all of the assessed domains

Table 3. Alternate-Form Reliability for WD.

	Fall			Winter			Spring		
WD metric	First	Second	Third	First	Second	Third	First	Second	Third
WW	.81**	.87**	.87**	.94**	.92**	.88**	.97**	.96**	.97**
WSC	.97**	.91**	.91**	.92**	.93**	.88**	.92**	.97**	.92**
CLS	.95**	.93**	.91**	.95**	.94**	.91**	.97**	.97**	.94**
ILS	.73**	.63**	.65**	.94**	.88**	.84**	.98**	.92**	.61**
%WSC	.91**	.79**	.82**	.83**	.92**	.89**	.88**	.95**	.87**
%CLS	.89**	.78**	.81**	.89**	.92**	.88**	.92**	.93**	.78**
C-ILS	.95**	.90**	.89**	.94**	.93**	.91**	.97**	.97**	.86**

Note. WD = word dictation; WW = words written; WSC = words spelled correctly; CLS = correct letter sequences; ILS = incorrect letter sequences; C-ILS = correct minus incorrect letter sequences.
*$p < .05$. **$p < .01$.

Table 4. Pearson's Correlations Between Fall and Winter WD With ACCESS-W.

Measure	Grade						
	1	2	3	4	5	6	7
			First grade (n = 21)				
1. ACCESS-W	—	.26	.57**	.56*	.52*	.47*	.51*
2. WD–WW	.43	—	.48*	.64**	.23	−.03	.12
3. WD–WSC	.65**	.75**	—	.92**	.94**	.71**	.82**
4. WD–CLS	.61**	.90**	.95**	—	.79**	.70**	.84**
5. WD–%WSC	.64**	.46*	.89**	.77**	—	.82**	.84**
6. WD–%CLS	.61**	.29	.73**	.63**	.88**	—	.94**
7. WD–C-ILS	.68**	.66**	.96**	.92**	.89**	.86**	—
			Second grade (n = 24)				
1. ACCESS-W	—	.62**	.81**	.78**	.70**	.60**	.75**
2. WD–WW	.63**	—	.71**	.87**	.32	.25	.65**
3. WD–WSC	.73**	.77**	—	.95**	.87**	.82**	.97**
4. WD–CLS	.72**	.89**	.97**	—	.71**	.68**	.94**
5. WD–%WSC	.65**	.47*	.90**	.80**	—	.95**	.86**
6. WD–%CLS	.54**	.42*	.85**	.77**	.96**	—	.88**
7. WD–C-ILS	.68**	.74**	.99**	.96**	.91**	.90**	—
			Third grade (n = 24)				
1. ACCESS-W	—	.63**	.69**	.68**	.54**	.49*	.66**
2. WD–WW	.64**	—	.70**	.89**	.32	.34	.68**
3. WD–WSC	.69**	.86**	—	.94**	.89**	.86**	.99**
4. WD–CLS	.68**	.96**	.96**	—	.70**	.70**	.94**
5. WD–%WSC	.60**	.57**	.89**	.76**	—	.97**	.89**
6. WD–%CLS	.54**	.57**	.84**	.75**	.95**	—	.89**
7. WD–C-ILS	.67**	.87**	.98**	.97**	.87**	.87**	—

Note. Intercorrelations for concurrent validity (winter) are presented above the diagonal and intercorrelations for predictive validity (fall) are presented below the diagonal. WD = word dictation; ACCESS-W = ACCESS-Writing; WW = average words written; WSC = average words spelled correctly; CLS = average correct letter sequences; %WSC = total WSC/total WW; %CLS = total CLS/total letter sequences; C-ILS = average CLS minus average incorrect letter sequences.
*$p < .05$. **$p < .01$.

(i.e., reading, listening, speaking; Salvia et al., 2017). Furthermore, a review of mean performance for each grade level at each benchmark indicates that student scores increased across the year, which provides preliminary evidence that future research examining WD Stage 2 research is warranted. If the participants in this study had not

Table 5. Pearson's Correlations Between Word Dictation Metrics and ACCESS Subtests of Writ, Read, Spk, and Lstn.

	First grade				Second grade				Third grade			
Time-metric	Writ	Read	Spk	Lstn	Writ	Read	Spk	Lstn	Writ	Read	Spk	Lstn
Fall-WSC	.65**	.63**	.41	.58**	.73**	.63**	.53*	.52*	.69**	.61**	.43*	.59**
Fall-CLS	.61**	.54**	.37	.57**	.72**	.62**	.51*	.52*	.68**	.62**	.47*	.61*
Fall-C-ILS	.68**	.67**	.40	.62**	.68**	.59**	.54**	.48	.67**	.59**	.49*	.63**
Win-WSC	.57**	.61**	.34	.53**	.81**	.61**	.52**	.34	.69**	.59**	.34	.47*
Win-CLS	.56**	.50*	.33	.55**	.78**	.53**	.43*	.31	.68**	.61**	.17	.43*
Win-C-ILS	.51*	.50*	.23	.52*	.75**	.55**	.55**	.32	.66**	.59**	.38	.51*

Note. Writ = writing; Read = reading; Spk = speaking; Lstn = listening; Win = winter; WSC = words spelled correctly; CLS = correct letter sequences; C-ILS = correct minus incorrect letter sequences.
*$p < .05$. **$p < .01$.

demonstrated growth across benchmarks (i.e., 10 weeks) then it would be unlikely that they or similar students would demonstrate growth across the shorter timeframes (i.e., 8 weeks) necessary for progress monitoring within RTI.

In general, CLS appears to be the most reliable and valid WD metric for ELs across first through third grades and has potential for the purposes of screening. However, both C-ILS and WSC also displayed potential and should be examined in future research. This study adds to the existing theoretical, intervention, and CBM-Writing research evidencing that transcriptions skills (i.e., spelling) as captured by WD are reliable and valid predictors of overall writing performance by extending it to include young ELs, especially ELs with beginning to intermediate ELP.

Limitations and Future Directions

Several limitations serve to caution the interpretation of results and the extent to which they may be generalized to the larger EL population. Socioeconomic data were not available at the individual student level and all participants in this study received English-only instruction; therefore, results should not be generalized to ELs receiving various bilingual instruction models, and the influence of socioeconomic status is unknown. Although several native languages were represented, low numbers of ELs speaking any specific native language other than Spanish did not allow for examining how WD performed according to an EL's native language. In addition, data regarding native language proficiency were not available. Therefore, teachers should consider the possible implications of their EL's native language and native language proficiency when implementing WD and interpreting their performance. Other limitations include a small sample size, lack of ELP data in the fall, and that the majority of ELs included in the study had beginning and intermediate ELP in the winter. Due to the small sample size, results from this study should be considered preliminary and exploratory in nature. Furthermore, it is impossible to say how initial ELP (fall) may have influenced WD performance. Replication across regions/school districts is needed to examine the extent to which findings from this study may be generalized to other ELs.

Implications for Practice and Conclusions

Screening data should only be one part of a multifaceted assessment process, especially for intensive intervention and special education referral (Fletcher & Miciak, 2019). Teachers should refer to Burr and colleagues (2015) regarding linguistically and culturally responsive evaluation processes for ELs in special education. It is reasonable to use WD to drive low-stakes educational decisions and to use a combination of WD static score, especially CLS, and progress monitoring data to inform decisions to provide supplemental intervention (i.e., Tier II) and to *inform* the referral process. However, in lieu of a lack of Stage 2 (i.e., sensitivity to growth) research for WD with young ELs, progress monitoring should be conducted using a variety of sources, such as teacher-generated assessment and writing samples evaluated with established rubrics, in addition to WD.

Considering that RTI has been identified as the most reliable and valid method for identifying LD (Fletcher & Miciak, 2019) and that ELs experience delayed access to early literacy support (NASEM, 2017), it is critical that the field initiates lines of Stages 2 and 3 research in WD for young ELs. Practitioners not only need to be able to identify potential risk early but also need guidance in determining whether or not an EL is responding adequately to instruction and how to use the data generated by WD to improve student outcomes. Specific areas of interest include how the technical adequacy of static scores and expected rates of growth change as a function of an EL's model of instruction and ELP. Furthermore, studies may also begin examining the utility of WD for older ELs with beginning ELP who may not yet have the prerequisite skills for passage-level CBM-Writing like story prompt.

Finally, although the guide by Graham and colleagues (2012) provides general guidelines for evidence-based early

writing instruction, there is a need for research examining and establishing evidence-based writing intervention specifically for ELs to use RTI as a method for LD identification. Furthermore, spelling is not likely to capture all ELs who are at risk for or have an LD in writing because LD in writing may manifest in one or more domains (i.e., text generation or executive functioning), especially as grade-level curricula increasingly demand proficiency in higher level writing skills (i.e., vocabulary, syntax, organization). Research should continue mapping CBM-Writing types according to grade-level curriculum expectations and possibly also with ELP for ELs, so that the most appropriate skills are being assessed and monitored as writers develop.

Authors' Note

This article was selected for the Outstanding Researcher Award by the Council for Learning Disabilities in the fall of 2019.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The research reported here was supported in part by the Institute of Education Sciences, U.S. Department of Education, through Grant R324A130144 to the University of Minnesota. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

References

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*, 281–298. https://doi.org/10.1037/a0019318

Abedi, J., & Gándara, P. (2006). Performance of English Language Learners as a subgroup in large-scale assessment: Interaction of research and policy. *Educational Measurement: Issues and Practice*, *25*(4), 36–46. https://doi.org/10.1111/j.1745-3992.2006.00077.x

Aud, S., Hussar, W., Johnson, F., Kena, G., Roth, E., Manning, E., . . . Zhang, J. (2012). *The condition of education 2012* (NCES 2012-045). National Center for Education Statistics.

Berninger, V. W., & Abbott, R. D. (2010). Listening comprehension, oral expression, reading comprehension, and written expression: Related yet unique language systems in grades 1, 3, 5, and 7. *Journal of Educational Psychology*, *102*, 635–651. https://doi.org/10.1037/a0019319

Berninger, V. W., Abbott, R. D., Abbott, S. P., Graham, S., & Richards, T. (2002). Writing and reading: Connections between language by hand and language by eye. *Journal of Learning Disabilities*, *35*(1), 39–56. https://doi.org/10.1177/002221940203500104

Berninger, V. W., & Amtmann, D. (2003). Preventing written expression disabilities through early and continuing assessment and intervention for handwriting and/or spelling problems: Research into practice. In H. Swanson, K. Harris, & S. Graham (Eds.), *Handbook of learning disabilities* (pp. 323–344). Guilford.

Berninger, V. W., Raskind, W., Richards, T., Abbott, R., & Stock, P. (2008). A multidisciplinary approach to understanding developmental dyslexia within working-memory architecture: Genotypes, phenotypes, brain, and instruction. *Developmental Neuropsychology*, *33*(6), 707–744. https://doi.org/10.1080/87565640802418662

Berninger, V. W., & Swanson, H. L. (1994). Modifying Hayes and Flower's model of skilled writing to explain beginning and developing writing. In E. C. Butterfield (Ed.), *Advances in cognition and educational practice. Vol. 2: Children's writing: Toward a process theory of the development of skilled writing* (pp. 57–81). JAI Press.

Berninger, V. W., Vaughan, K. B., Abbott, R. D., Abbott, S. P., Rogan, L. W., Brooks, A., . . . & Graham, S. (1997). Treatment of handwriting problems in beginning writers: Transfer from handwriting to composition. *Journal of Educational Psychology*, *89*(4), 652. https://doi.org/10.1037/0022-0663.89.4.652

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

Burr, E., Haas, E., & Ferriere, K. (2015). *Identifying and supporting English learner students with learning disabilities: Key issues in the literature and state practice* (REL 2015-086). Regional Educational Laboratory West.

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

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

Costa, L. J. C., Edwards, C. N., & Hooper, S. R. (2016). Writing disabilities and reading disabilities in elementary school students: Rates of co-occurrence and cognitive burden. *Learning Disability Quarterly*, *39*(1), 17–30. https://doi.org./10.1177/0731948714565461

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

Espin, C., Wallace, T., Campbell, H., Lembke, E. S., Long, J. D., & Ticha, R. (2008). Curriculum-based measurement in writing: Predicting the success of high-school students on state standards tests. *Exceptional Children*, *74*(2), 174–193. https://doi.org/10.1177/001440290807400203

Fletcher, J. M., & Miciak, J. (2019). *The identification of specific learning disabilities: A summary of research on best practices*. Texas Center for Learning Disabilities.

Fox, J., & Fairbairn, S. (2011). Test review: ACCESS for ELLs [R]. *Language Testing*, *28*(3), 425–431. https://doi.org/10.1177/0265532211404195

Fry, R. (2010). *Hispanics, high school dropouts and the GED*. Pew Hispanic Center.

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

Fuchs, L. S., Deno, S. L., & Mirkin, P. K. (1984). The effects of frequent curriculum-based measurement and evaluation on pedagogy, student achievement, and student awareness of learn-in. *American Educational Research Journal*, *21*, 449–460. https://doi.org/10.3102/00028312021002449

Genesee, F., Lindholm-Leary, K., Saunders, W., & Christian, D. (2005). English language learners in US schools: An overview of research findings. *Journal of Education for Students Placed at Risk*, *10*(4), 363–385. https://doi.org/10.1207/s15327671espr1004_2

Ghasemi, A., & Zahediasl, S. (2012). Normality tests for statistical analysis: A guide for non-statisticians. *International Journal of Endocrinology and Metabolism*, *10*(2), 486–489. https://doi.org/10.5812/ijem.3505

Graham, S., Berninger, V. W., Abbott, R. D., Abbott, S. P., & Whitaker, D. (1997). Role of mechanics in composing of elementary school students: A new methodological approach. *Journal of Educational Psychology*, *89*(1), 170–182. https://doi.org./10.1037/0022-0663.89.1.170

Graham, S., Bolinger, A., Booth Olson, C., D'Aoust, C., MacArthur, C., McCutchen, D., & Olinghouse, N. (2012). *Teaching elementary school students to be effective writers: A practice guide* (NCEE 2012-4058). http://ies.ed.gov/ncee/wwc/publications_reviews.aspx#pubsearch

Graham, S., Harris, K. R., & Chorzempa, B. F. (2002). Contribution of spelling instruction to the spelling, writing, and reading of poor spellers. *Journal of Educational Psychology*, *94*(4), 669–686. https://doi.org/10.1037/0022-0663.94.4.669

Graham, S., Harris, K. R., & Larsen, L. (2001). Prevention and intervention of writing difficulties for students with learning disabilities. *Learning Disabilities Research & Practice*, *16*(2), 74–84. https://doi.org/10.1111/0938-8982.00009

Graham, S., & Hebert, M. (2011). Writing to read: A meta-analysis of the impact of writing and writing instruction on reading. *Harvard Educational Review*, *81*(4), 710–744. https://doi.org/10.17763/haer.81.4.t2k0m13756113566

Graham, S., & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology*, *99*(3), 445–476. https://doi.org/10.1037/0022-0663.99.3.445

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

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

Jones, D., & Christensen, C. (1999). The relationship between automaticity in handwriting and students' ability to generate written text. *Journal of Educational Psychology*, *91*, 44–49. https://doi.org/10.1037/0022-0663.91.1.44

Katusic, S. K., Colligan, R. C., Weaver, A. L., & Barbaresi, W. J. (2009). The forgotten learning disability: Epidemiology of written-language disorder in a population-based birth cohort (1976–1982). *Pediatrics*, *123*(5), 1306–1313. https://doi.org/10.1542/peds.2008-2098

Keller-Margulis, M. A., Clemens, N. H., Im, M. H., Kwok, O. M., & Booth, C. (2012). Curriculum-based measurement yearly growth rates: An examination of English language learners and native English speakers. *Learning and Individual Differences*, *22*(6), 799–805. https://doi.org/10.1016/j.lindif.2012.07.005

Keller-Margulis, M. A., Ochs, S., Reid, E. K., Faith, E. L., & Schanding, G. T. (2019). Validity and diagnostic accuracy of early written expression screeners in kindergarten. *Journal of Psychoeducational Assessment*, *37*(5), 539–552. https://doi.org/10.1177/0734282918769978

Keller-Margulis, M. A., 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

Kena, G., Hussar, W., McFarland, J., de Brey, C., Musu-Gillette, L., Wang, X., Zhang, J., Rathbun, A., Wilkinson-Flicker, S., Diliberti, M., Barmer, A., Bullock Mann, F., & Dunlop Velez, E. (2016). *The condition of education 2016* (NCES 2016-144). http://nces.ed.gov/pubsearch

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

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

McMaster, K. L., & Espin, C. (2007). Technical features of curriculum-based measurement in writing a literature review. *The Journal of Special Education*, *41*(2), 68–84. https://doi.org/10.1177/00224669070410020301

McMaster, K. L., & Lembke, E. S. (2014). *Data-based instruction for beginning writing: Tools, learning, and coaching*. Department of Educational Psychology, University of Minnesota.

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

McMaster, K. L., Ritchey, K. D., & Lembke, E. S. (2011). Curriculum-based measurement for beginning writers: Recent developments and future directions. In T. E. Scruggs & M. A. Mastropieri (Eds.), *Assessment and intervention: Advances in learning and behavioral disabilities* (*Vol. 24*, pp. 111–148). Emerald. https://doi.org/10.1108/S0735-004X(2011)0000024008

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–104). Macmillan.

Miller, B., & McCardle, P. (2011). Reflections on the need for continued research on writing. *Reading & Writing*, *24*(2), 121–132. https://doi.org/10.1007/s11145-010-9267-6

Mueller, P. A., & Oppenheimer, D. M. (2014). The pen is mightier than the keyboard: Advantages of longhand over laptop note taking. *Psychological Science*, *25*, 1159–1168. https://doi.org/10.1177/0956797614524581

National Academies of Sciences, Engineering, and Medicine. (2017). *Promoting the education success of children and youth learning English: Promising futures*. The National Academies Press. https://doi.org./10.17226/24677

National Center for Education Statistics. (2012). *Data explorer*. http://nces.ed.gov/nationsreportcard/naepdata/dataset.aspx

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

Parker, C. E., Louie, J., & O'Dwyer, L. (2009). *New measures of English language proficiency and their relationship to performance on large-scale content assessments* (Issues & Answers Report, REL 2009-No. 066). Regional Educational Laboratory.

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

Ritchey, K. D., & Coker, D. L. (2014). Identifying writing difficulties in first grade: An investigation of writing and reading measures. *Learning Disabilities Research & Practice*, *29*(2), 54–65. https://doi.org/10.1111/ldrp.12030

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

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

Taylor, R. L. (2006). *Assessment of exceptional students: Educational and psychological procedures* (7th ed.). Pearson Education.

Yanosky, T., Yen, S. J., Louguit, M., MacGregor, D., Zhang, Y., & Kenyon, D. (2011). *Annual technical report for ACCESS for ELLs® English language proficiency test, series 200, 2008-2009 administration* (WIDA Consortium Annual Technical Report No. 6). The University of Wisconsin.