Journal for Educational Research Online
Journal für Bildungsforschung Online
https://doi.org/10.31244/jero.2022.01.03
Volume 14 (2022), No. 1, 39–60

Sterett H. Mercer & Joanna E. Cannon

Validity of Automated Learning Progress Assessment in English Written Expression for Students with Learning Difficulties

Abstract

We evaluated the validity of an automated approach to learning progress assessment (aLPA) for English written expression. Participants (n = 105) were students in Grades 2–12 who had parent-identified learning difficulties and received academic tutoring through a community-based organization. Participants completed narrative writing samples in the fall and spring of 1 academic year, and some participants (n = 33) also completed a standardized writing assessment in the spring of the academic year. The narrative writing samples were evaluated using aLPA, four hand-scored written expression curriculum-based measures (WE-CBM), and ratings of writing quality. Results indicated (a) aLPA and WE-CBM scores were highly correlated with ratings of writing quality; (b) aLPA and more complex WE-CBM scores demonstrated acceptable correlations with the standardized writing subtest assessing spelling and grammar, but not the subtest assessing substantive quality; and (c) aLPA scores showed small, statistically significant improvements from fall to spring. These findings provide preliminary evidence that aLPA can be used to efficiently score narrative writing samples for progress monitoring, with some evidence that the aLPA scores can serve as a general indicator of writing skill. The use of automated scoring in aLPA, with performance comparable to WE-CBM hand scoring, may improve scoring feasibility and increase the likelihood that educators implement aLPA for decision-making.

Keywords

automated text evaluation, learning progress assessment, written expression, curriculum-based measurement, learning difficulties

Prof. Dr. Sterett H. Mercer (corresponding author), ORCID: 0000-0002-7940-4221 · Prof. Dr. Joanna E. Cannon, ORCID: 0000-0003-4552-4476, Department of Educational and Counselling Psychology and Special Education, University of British Columbia, 2125 Main Mall, Vancouver, BC V6T 1Z4, Canada
email: sterett.mercer@ubc.ca
joanna.cannon@ubc.ca

Validität des automatisierten Learning Progress Assessments im geschriebenen Englisch für Schüler:innen mit Lernschwierigkeiten

Zusammenfassung

Wir evaluierten die Validität eines automatischen Ansatzes des Learning Progress Assessments (aLPA) in geschriebener Sprache. Schüler:innen der Klassen 2 bis 12 (n = 105) mit Lernschwierigkeiten, die deren Eltern festgestellt hatten, nahmen an Nachhilfeunterricht, welcher von einer gemeinnützigen Organisation durchgeführt wurde, teil. Die Schüler:innen erstellten im Herbst und Frühjahr eines Schuljahres Schreibproben. Weiterhin nahmen einige Schüler:innen (n = 33) an einer standardisierten Schreibprüfung im Frühjahr teil. Die Schreibproben wurden mit aLPA, vier handkodierten Curriculum-Based Measures (WE-CBM) und hinsichtlich deren Schreibqualität ausgewertet. Unsere Ergebnisse zeigen, (a) aLPA- und WE-CBM-Werte korrelierten hoch mit der Bewertung der Schreibqualität, (b) aLPA- und die komplexeren WE-CBM-Werte zeigten akzeptable Korrelationen mit der standardisierten Rechtschreib- und Grammatikprüfung, jedoch nicht mit tatsächlicher Qualität, und (c) aLPA-Werte zeigten geringe, statistisch signifikante Verbesserungen von Herbst zu Frühjahr. Diese Ergebnisse deuten darauf hin, dass aLPA als effiziente Methode zur Bewertung von Schreibproben verwendet werden kann und der aLPA-Wert als allgemeiner Schreibkompetenzindikator dienen kann. Automatische Bewertung im aLPA, dessen Validität mit der von WE-CBM vergleichbar ist, kann die Bewertung vereinfachen und damit ist es wahrscheinlicher, dass Lehrkräfte aLPA verwenden.

Schlagworte

Lernverlaufsdiagnostik, automatisiertes Learning Progress Assessment, geschriebenes Englisch, Schüler:innen mit Lernschwierigkeiten

1. Introduction

Many students with academic difficulties do not respond to generally effective academic interventions (McMaster et al., 2005); this failure to respond is problematic because educators and related professionals are responsible for improving the academic skills of individual students. To address this concern, data-based individualization is gaining recognition as an approach to special education service delivery (DBI; Fuchs et al., 2013). In DBI, generally effective interventions, as determined by the research literature, are first implemented with ongoing progress monitoring of student academic outcomes. When progress monitoring data indicate that the intervention is not working for a specific student, educators modify the intervention approach until monitoring data indicate improvement. Although DBI is commonly implemented and researched for reading interventions (Filder-

man et al., 2018), DBI in written expression is less commonly implemented (Jung et al., 2018), in part due to challenges related to progress monitoring assessment. The purpose of this study is to present a validity argument and preliminary validity evidence for automated learning progress assessment (aLPA) in written expression.

1.1 Curriculum-Based Measurement

In DBI, progress monitoring assessment is typically done using curriculum-based measurement (CBM; Deno, 1985). Several characteristics are emphasized in CBM: (a) simplicity and efficiency of administration, scoring, and interpretation; (b) general outcome measurement, meaning that these simple measures should serve as indicators of overall performance in an academic domain (e.g., correlate with more comprehensive assessments of academic skill); (c) cost-effectiveness so that alternate forms can be frequently administered; and (d) reliability and validity of scores to inform data-based instructional decisions. Although CBM originally used curriculum-specific materials in assessment, the emphasis on reliability and validity led to the development of standard CBM tasks and materials that are aligned to the general skills taught in the curriculum (Hosp et al., 2016). Oral passage reading (OPR) is a commonly-used CBM that consists of a 1-minute timed reading of a standardized, field-tested, grade-level passage that is scored for the number of words read correctly. Despite the brevity and simplicity, OPR scores correlate highly with more comprehensive standardized assessments of reading (Reschly et al., 2009) and are sensitive to differential skill growth during intervention (Morgan & Sideridis, 2006).

For written expression CBM (WE-CBM), typical procedures involve presenting students with a short story starter (e.g., “One day on the way to school, I …”), allowing them to plan for 1 minute, and then having them write for 3 minutes (Hosp et al., 2016). These writing samples are then hand-scored with simple metrics such as counts of the total number of words written and the number of words spelled correctly or with more complex metrics such as the number of correct word sequences (the number of grammatically and semantically acceptable adjacent words that are spelled and punctuated correctly; Videen et al., 1982). Although there is some research to support the use of these WE-CBM procedures as part of DBI to improve elementary students’ early writing skills (McMaster et al., 2020), there are several challenges to be addressed at higher grade levels, including the need for multiple, longer duration writing samples for adequate reliability (Keller-Margulis et al., 2016; Kim et al., 2017) and the requirement to use more complex scoring metrics for adequate validity (Mercer et al., 2012; Romig et al., 2017). Complex WE-CBM scoring of longer duration writing samples may limit feasibility of implementation by educators (Espin et al., 1999), which could be a barrier to more widespread implementation of DBI for written expression beyond the early elementary grades.

1.2 Automated Learning Progress Assessment in Written Expression

To address these concerns, we describe an automated approach to learning progress assessment as an alternative to traditional hand-scored WE-CBM. We describe our approach using the more general term LPA, consistent with Förster and Souvignier's (2014, 2015) LPA in reading, for several reasons. First, we believe that scoring methods for written expression, although summarized in an overall writing skill score, should be based on a comprehensive set of writing quality indices (an extended test concept; Förster & Souvignier, 2015). In WE-CBM, scoring primarily considers text production and accuracy, whereas in aLPA, we use a wide range of word-, sentence-, and discourse-level indices provided by automated text evaluation software to generate overall writing quality scores. Second, based on research demonstrating that automated text evaluation can generate writing quality scores that are useful for screening (Keller-Margulis et al., 2021; Mercer et al., 2019; Wilson, 2018), we also anticipate that computer-based assessment will be necessary for writing samples to be scored and for such a system to be feasibly used by teachers. Third, given that we know that multiple, longer-duration writing samples will be necessary for reliability (Keller-Margulis et al., 2016), we anticipate that a reduced test frequency will be optimal, compared to typical CBM progress monitoring procedures of weekly assessments. We differ from Förster and Souvignier (2014, 2015) in our primary intended uses for aLPA, specifically, screening and progress monitoring of students with learning difficulties instead of monitoring all students, but concur that such a system should also be useful for monitoring the learning progress of most students.

1.2.1 Interpretation and Use of aLPA

Consistent with contemporary validation approaches (American Educational Research Association et al., 2014; Kane, 2013), we specify our proposed interpretation and use of aLPA scores, and also continue to contrast aLPA with WE-CBM. Specifically, we believe that aLPA scores should assess the overall writing quality of student writing samples generated under authentic instructional conditions. By authentic, we mean, at minimum, that students should have adequate time to engage in the writing process (i.e., beyond the 3-minute time limit often used in WE-CBM), and ideally the writing samples would be generated from tasks embedded in the curriculum, "where purpose and audience are clear and meaningful, where support and feedback are readily available, and where the final product has academic value for the student" (Calfee & Miller, 2007, p. 269).

1.2.1.1 Scoring Inference

In stating that aLPA scores should reflect writing quality on the evaluated writing samples, we are articulating an explicit, testable *scoring inference* (Kane, 2013) that is informed by developmental writing theory. The not-so-simple view of writing (Berninger & Amtmann, 2003) considers writing to involve three processes that are constrained by working memory capacity: (a) text generation, which involves idea generation and translation of these ideas into language; (b) transcription skills, which is the translation of these language representations into written text by handwriting or typing; and (c) strategic self-regulation of the writing process (e.g., planning, drafting, revising). Difficulties in text generation or transcription skills can contribute to poor writing fluency, and this limited automaticity in generating connected text can impair writing quality (Kim et al., 2018). For developing writers, writing quality is often defined as the extent to which ideas are developed and organized in compositions (Kim et al., 2015, 2018).

Due to short time allowed to write, WE-CBM scores are typically interpreted as assessing the construct of writing fluency more so than writing quality (Kim et al., 2018; Ritchey et al., 2016), and the extent to which WE-CBM scores assess quality on the writing samples used to generate the scores has received minimal attention in research (McMaster & Espin, 2007). By contrast, in aLPA, students have a more extended time to write, thereby increasing students' ability to engage in the processes described in the not-so-simple view of writing. Also, we explicitly test the scoring inference that automated aLPA scores assess raters' judgements of quality on the evaluated writing samples.

1.2.1.2 Generalization Inference

In addition to assessing quality on the administered writing samples, aLPA scores should also correlate with quality on writing samples administered under similar conditions; this claim is a *generalization inference* (Kane, 2013). Because a generalizability theory study of WE-CBM scores demonstrated that reliability improves as writing sample duration increases (durations of 1–7 minutes were investigated; Keller-Margulis et al., 2016), we anticipate that the longer administration time in aLPA will support generalization. That said, a longer duration may not be sufficient given that generalizability theory studies have found that multiple writing samples are needed for adequate reliability and there may be limited generalization across writing samples of different genres (Kim et al., 2017; Wilson et al., 2019).

1.2.1.3 Extrapolation Inference

Similar to WE-CBM, we expect that aLPA scores will serve as general indicators of writing skill, which is an *extrapolation inference* (Kane, 2013). This inference is

testable by determining the extent to which aLPA scores correlate with scores on other standardized writing assessments that have evidence of validity. In CBM and aLPA, scores are used for screening and progress monitoring; because these decisions involve the assessment of large groups of students (screening) or frequent assessment of skills (progress monitoring), efficiency of administration and scoring are important (Deno, 1985). For this reason, evidence of the ability to extrapolate is critical – the goal in CBM and aLPA is to identify the most efficient test format that can maintain strong correlations with more comprehensive assessments of academic skills.

1.2.1.4 Decision Inference

Several sources of evidence can support the validity of aLPA for screening and progress monitoring, which are *decision inferences* (Kane, 2013). For example, screening instruments should have evidence that scores can be used to accurately predict meaningful outcomes such as placements of students in special education programs (e.g., Fewster & Macmillan, 2002) and whether students meet proficiency standards on high-stakes assessments (e.g., Furey et al., 2016). Progress monitoring instruments should have evidence that scores improve in response to writing intervention (e.g., McMaster et al., 2017). Ultimately, a set of decision rules would need to be articulated and evaluated to use aLPA for screening and progress monitoring, which would require the establishment of aLPA norms and performance standards.

1.3 Current Study

As an early-stage investigation of the validity of aLPA for progress monitoring in written expression, we compare the performance of aLPA vs. WE-CBM metrics in relation to some of the previously detailed assumptions related to scoring, extrapolation, and decision inferences. Analyses are based on writing samples from students with parent-reported learning difficulties receiving academic tutoring in a community agency for 1 academic year. Specifically, we address the following research questions:

1. Scoring inference: How strongly do aLPA quality scores and WE-CBM metrics correlate with ratings of writing quality on the scored writing samples?
2. Extrapolation inference: How strongly do aLPA quality scores and WE-CBM metrics correlate with scores on a standardized writing assessment?
3. Decision inference: Are aLPA quality scores sensitive to student skill growth?

2. Method

2.1 Participants and Setting

All participants were enrolled in a community-based nonprofit organization that provides low-cost, after-school, one-to-one academic tutoring for approximately 2 hours per week in reading, math, and/or written expression to students with suspected or diagnosed learning disabilities. As part of typical service delivery, the organization collected picture-prompted narrative writing samples from students in the fall (September–October) and spring (April–May) of an academic year to inform instruction, and some students also completed a standardized writing assessment in the spring. Participants included 105 students in Grades 2–12 who completed at least one of the narrative writing samples; 79 participants completed both the fall and spring writing samples. Of the participants, 103 completed the fall writing sample, 83 completed the spring writing sample, and 33 completed the standardized writing assessment. The participants were in the following grade levels in school: Grade 2 (n = 11), 3 (n = 20), 4 (n = 14), 5 (n = 20), 6 (n = 13), 7 (n = 8), 8 (n = 6), 9 (n = 7), 10 (n = 2), 11 (n = 2), and 12 (n = 2). Forty percent of the participants were female. The 33 participants with standardized writing assessment data were in Grades 3–9; 39% were female.

Because we only had access to the writing samples and assessment data, we do not have detailed demographic information for participants or the specific learning difficulties of students; however, all were experiencing academic difficulties substantial enough for parents to seek community-based tutoring beyond the school-based supports available. Formal diagnoses of learning disabilities are not required by the organization due to long wait lists for assessments in the local public schools and high costs for assessments in private clinics (Werb, 2007). Approximately 50% of students served by the organization have a formal diagnosis of a learning disability, and approximately 60% of students receive financial support to access the organization's services.

Most participants attended school in a large, urban, ethnically diverse school district of approximately 52 000 students in Western Canada. In the district, 44% of students speak a language other than English at home, with 160 different languages spoken by families in the district. The top five languages other than English spoken at home are the following: Cantonese (17% of students), Mandarin (11%), Tagalog (5%), Vietnamese (4%), and Punjabi (4%). Approximately 17% of students in the district are eligible for English language supports, and 11% of students are eligible for special education services.

2.2 Measures

Measures included ratings of writing quality, aLPA quality, and multiple WE-CBM metrics based on narrative writing samples and selected subtests from a standardized writing assessment.

2.2.1 Narrative Writing Samples

To inform instruction, the non-profit agency's tutors asked students to select one picture about which they would like to write from a collection of photos from travel, recreation, and lifestyle magazines (e.g., pictures of amusement park rides, animals, restaurants). The tutors informed students that they would have 10 minutes to handwrite, and the tutors provided no assistance to students during this time period. As part of their typical planning and assessment model, tutors recorded on a checklist whether the students' writing exhibited specific characteristics consistent with grade-level expectations, for example, subject-verb agreement, capitalization, punctuation, and organization; because the current study focused on overall writing quality, we did not use the checklist data in the current study. Before scoring by hand (rated writing quality and WE-CBM) and computer (aLPA), a research team member typed all writing samples, preserving errors in spelling and grammar, and a second research team member verified the accuracy of all transcriptions.

2.2.1.1 Rated Writing Quality

We hand-evaluated writing quality for the picture-prompted writing samples using the method of paired comparisons (Thurstone, 1927). Specifically, two raters each completed 3000 comparisons of pairs of the 186 writing samples (103 from the beginning and 83 from the end of the academic year). In these comparisons, raters identified the writing sample in the pair that was of better overall quality considering idea development and organization; for writing samples of similar quality, raters were instructed to select the writing sample that they would prefer to continue reading if it were longer. Our evaluation of writing quality based on idea development (which composition has more detailed and rich ideas from unique or interesting perspectives?) and organization of ideas (which composition has a more logical sequence, with a beginning, middle, and end, and better transitioning?) is consistent with prior research on developing writers (e.g., Kim et al., 2015, 2018).

We converted the paired comparison quality data to Elo ratings (see Pelánek, 2016, for a description) using the EloChoice R package (Neumann, 2019); the ratings represent the likelihood that writing samples would be rated as of higher quality in additional paired comparisons. The strong correlation between the Elo ratings from the two raters (r = .94) and an index of the proportion of comparisons

in which the writing sample with the highest prior Elo rating was rated as of better quality (.87; see Clark et al., 2018, for computational details) provide evidence for interrater reliability and the stability of the ratings.

2.2.1.2 aLPA Writing Quality

We used the writeAlizer R package (Mercer, 2020), which applies an ensemble scoring model to indices generated from the open-source ReaderBench text complexity analysis tool (Dascalu et al., 2018), to generate aLPA writing quality scores. The writeAlizer scoring model was originally trained based on 7-minute narrative writing samples from students in Grades 2–5 (see Mercer et al., 2019). In these analyses, seven machine learning algorithms (random forest regression, cubist regression, support vector machines with a radial kernel, elastic net regression, bagged multivariate adaptive regression splines, stochastic gradient boosted trees, and partial least squares regression; for details on these algorithms, see Hastie et al., 2009) were differentially weighted to predict writing quality ratings (i.e., Elo ratings similar to the ones used in the current study). In addition to a narrower grade range than in the current study (Grades 2–5 vs. Grades 2–12), only 6% of participants in the writeAlizer training set were receiving special education services, compared to the current participants with parent-reported learning difficulties. More details of scoring model development, including the weightings of ReaderBench indices in each algorithm and overall, are available in the writeAlizer online documentation: https://github.com/shmercer/writeAlizer. To facilitate interpretability, aLPA scores, across the fall and spring time points, were standardized before analyses.

2.2.1.3 WE-CBM Scoring

We used the guidelines of Hosp et al. (2016) to hand-score four WE-CBM metrics: total words written (TWW), words spelled correctly (WSC), correct word sequences (CWS), and correct minus incorrect word sequences (CIWS). TWW are the number of letters or groups of letters that are separated by spaces, even if the words are misspelled or used incorrectly in context. WSC are the number of words spelled correctly not considering context; words are counted as correct if they appear in the English language. CWS are the number of adjacent words that are acceptable in the English language considering spelling, punctuation, syntax, and semantics. CIWS is calculated from the number of CWS minus incorrect word sequences. Forty percent of the writing samples were independently scored by two raters – agreement was high (r = .99) for TWW, WSC, CWS, and CIWS scores.

2.2.2 Standardized Writing Assessment

Two subtests of the Test of Written Language, 4th edition (TOWL-4; Hammill & Larsen, 2009), requiring students to generate one picture-prompted narrative sample (5 minutes to plan, 15 minutes to write), were administered by the non-profit organization's staff and were scored by the research team. Scoring for contextual conventions considers spelling and grammatical errors, and scoring for story composition considers quality of vocabulary, plot, and interest to the reader. The TOWL-4 is commonly used as a criterion measure in validity studies for WE-CBM (Romig et al., 2017).

2.3 Data Analysis

To address Research Questions 1 and 2 that involve the relations of scores with rated quality on the scored writing samples and with standardized writing assessment scores, we calculated Pearson *r* correlation coefficients. Consistent with WE-CBM validity studies, we interpreted correlation coefficients of $r = .50$ as the minimally sufficient value for validity evidence (McMaster & Campbell, 2008), with values above $r = .60$ preferred and consistent with evidence standards used by the National Center on Intensive Intervention (2018) for academic screening tools. Differences between aLPA and WE-CBM validity coefficients were evaluated with Meng et al.'s (1992) *z* test for dependent correlation coefficients using the cocor package (Diedenhofen & Musch, 2015), with Bonferroni-adjusted *p* values reported to address multiple comparisons within the same time point and criterion measure. To address Research Question 3, we used a paired sample *t* test to determine if mean aLPA scores changed from fall to spring, with Hedges' *g* reported as an indicator of effect size.

3. Results

Means and standard deviations for all scores by time point are presented in Table 1. aLPA scores were highly correlated with WE-CBM scores at both time points (see Table 2); aLPA scores were most strongly related to TWW, WSC, and CWS scores (fall: $r = .91–.94$; spring: $r = .88–.90$), and also had large magnitude correlations with CIWS scores (fall: $r = .74$; spring: $r = .76$). Results are presented below by research question.

Table 1: Means, Standard Deviations, and Sample Sizes for all Scores by Time Point

	Fall			Spring		
	M	*SD*	*n*	*M*	*SD*	*n*
aLPA quality	−0.13	1.00	103	0.16	0.98	83
TWW	58.03	40.07	103	71.93	47.20	83
WSC	50.49	40.25	103	64.80	46.72	83
CWS	38.37	36.03	103	53.95	47.53	83
CIWS	12.77	35.59	103	28.81	51.16	83
Rated quality	−0.15	1.00	103	0.18	0.96	83
TOWL-4: CC	9.67	5.03	33	10.00	5.15	30
TOWL-4: SC	7.21	3.63	33	7.40	3.71	30

Note. aLPA = automated Learning Progress Assessment; TWW = Total Words Written; WSC = Words Spelled Correctly; CWS = Correct Word Sequences; CIWS = Correct Minus Incorrect Word Sequences; TOWL-4 = Test of Written Language (4th ed.); CC = Contextual Conventions; SC = Story Composition. Although the TOWL-4 was administered only once at the spring time point, values are reported at both time points based on participants who completed both the narrative writing sample at that time point and the TOWL-4.

Table 2: Correlations of aLPA Quality With WE-CBM Scores

	Fall (n = 103)			Spring (n = 83)		
Score	*r*	95% CI for *r*		*r*	95% CI for *r*	
TWW	.91	.87	.94	.88	.81	.92
WSC	.94	.91	.96	.90	.85	.93
CWS	.93	.90	.95	.89	.83	.93
CIWS	.74	.63	.81	.76	.65	.84

Note. TWW = Total Words Written; WSC = Words Spelled Correctly; CWS = Correct Word Sequences; CIWS = Correct Minus Incorrect Word Sequences.

3.1 Correlations With Rated Quality

Correlations of rated quality on the scored writing samples with aLPA quality scores and WE-CBM scores in the fall and spring of the academic year are presented in Table 3. Also, a scatterplot of the relation between aLPA and rated quality scores at the fall time point is presented in Figure 1. In general, correlations with rated quality were of large magnitude (all rs $\geq$ .79), with aLPA demonstrating higher correlations than all WE-CBM metrics at both time points, although the differences between the aLPA and WE-CBM correlation coefficients were of small magnitude and not always statistically significant. In the fall, the correlation of rated quality with aLPA quality (r = .93) was significantly greater than the correlation of rated quality with TWW (r = .81; Δr = .12, z = 6.34, p < .001), WSC (r =

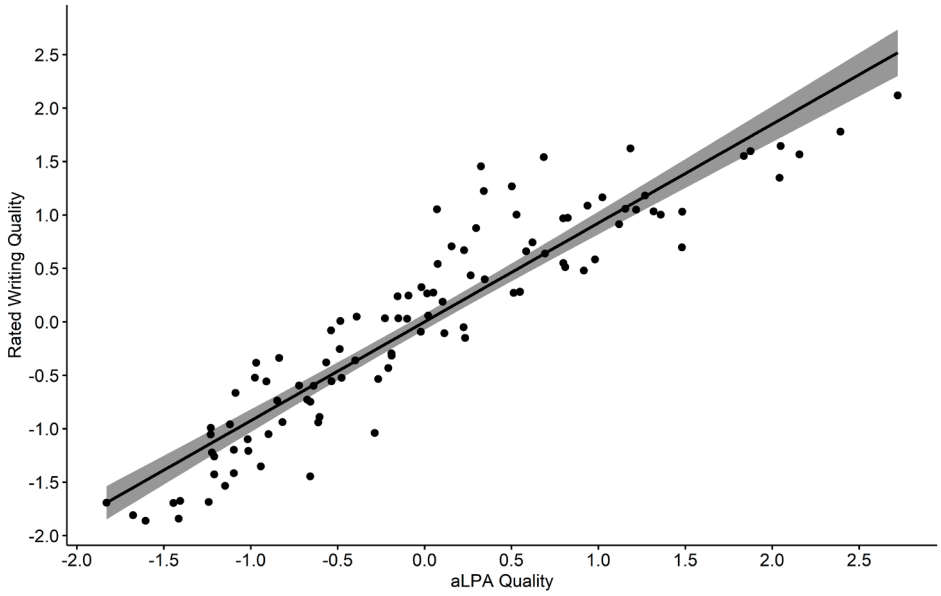
.86; Δr = .07, z = 4.43, p < .001), and CIWS (r = .79; Δr = .14, z = 4.95, p < .001), but did not statistically differ from the correlation of rated quality with CWS (r = .89; Δr = .04, z = 2.29, p = .111). In the spring, the correlation of rated quality with aLPA quality (r = .88) was significantly greater than the correlation of rated quality with TWW (r = .80; Δr = .08, z = 3.16, p = .006), but was not statistically different from the correlations of rated quality with WSC (r = .85; Δr = .03, z = 1.33, p = .739), CWS (r = .86; Δr = .02, z = 0.75, p = .999), or CIWS (r = .80; Δr = .08, z = 2.29, p = .089).

Table 3: Correlations With Rated Writing Quality by Time Point

	Fall (n = 103)			Spring (n = 83)		
Score	r	95% CI for r		r	95% CI for r	
aLPA quality	.93	.89	.95	.88	.82	.92
TWW	.81	.73	.87	.80	.70	.86
WSC	.86	.81	.90	.85	.78	.90
CWS	.89	.85	.93	.86	.80	.91
CIWS	.79	.70	.85	.80	.71	.87

Note. aLPA = automated Learning Progress Assessment; TWW = Total Words Written; WSC = Words Spelled Correctly; CWS = Correct Word Sequences; CIWS = Correct Minus Incorrect Word Sequences.

Figure 1: Relation of aLPA and Rated Writing Quality at the Fall Time Point



Note. n = 103. aLPA and rated quality scores are standardized across the fall and spring time points (M = 0, SD = 1). The regression line is displayed as a solid line, with shading indicating its 95% confidence interval.

3.2 Correlations With a Standardized Writing Assessment

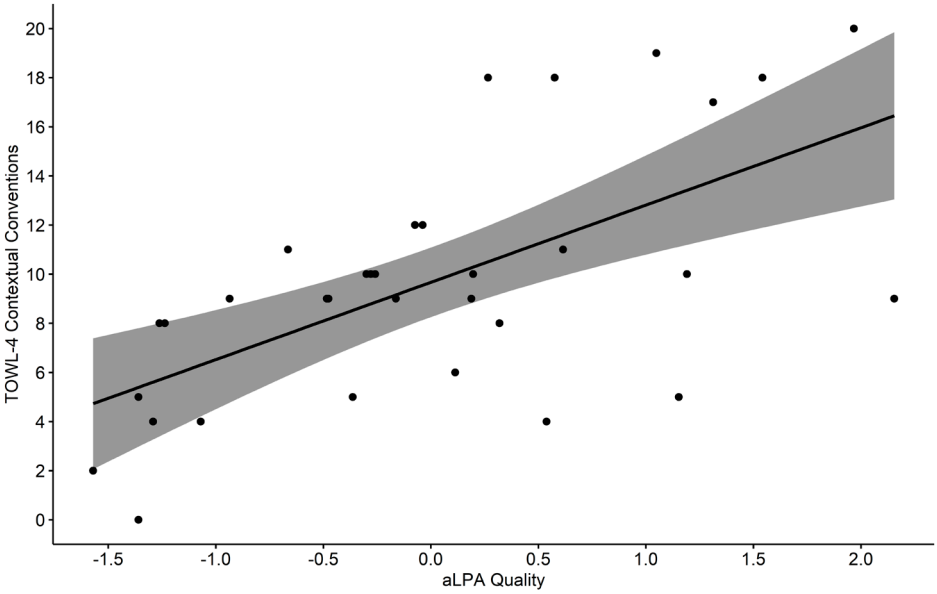
Validity coefficients in relation to the TOWL-4 subtest that assessed spelling and grammar (contextual conventions) and the subtest that assessed substantive quality (story composition) are presented below and in Table 4. Scatterplots of the relations between aLPA (fall time point) and the TOWL-4 subtest scores are presented in Figures 2 and 3.

Table 4: Correlations With a Standardized Writing Assessment

	Fall (n = 33)			Spring (n = 30)		
Score	r	95% CI for r		r	95% CI for r	
TOWL-4 contextual conventions						
aLPA quality	.63	.36	.80	.72	.49	.86
TWW	.48	.16	.70	.59	.29	.78
WSC	.54	.24	.75	.65	.38	.82
CWS	.66	.42	.82	.68	.42	.84
CIWS	.67	.43	.83	.67	.41	.83
TOWL-4 story composition						
aLPA quality	.44	.11	.68	.53	.20	.75
TWW	.34	.00	.61	.44	.10	.69
WSC	.36	.02	.63	.46	.12	.70
CWS	.45	.12	.69	.45	.10	.69
CIWS	.43	.10	.67	.38	.02	.65

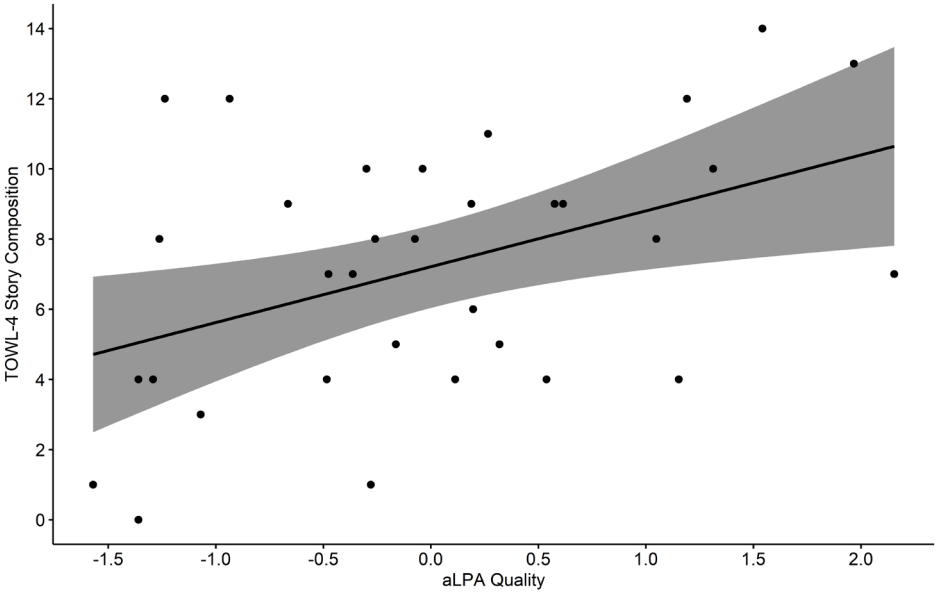
Note. TOWL-4 = Test of Written Language (4th ed.); aLPA = automated Learning Progress Assessment; TWW = Total Words Written; WSC = Words Spelled Correctly; CWS = Correct Word Sequences; CIWS = Correct Minus Incorrect Word Sequences.

Figure 2: Relation of aLPA Quality and TOWL-4 Contextual Conventions at the Fall Time Point



Note. n = 33. aLPA quality scores are standardized across the fall and spring time points (M = 0, SD = 1). The regression line is displayed as a solid line, with shading indicating its 95% confidence interval.

Figure 3: Relation of aLPA Quality and TOWL-4 Story Composition at the Fall Time Point



The regression line is displayed as a solid line, with shading indicating its 95% confidence interval.

3.2.1 Contextual Conventions

For the fall time point, aLPA (r = .63), CWS (r = .66), and CIWS (r = .67) had correlations above the r = .60 standard for criterion-related validity coefficients. The validity coefficient for aLPA not statistically different from the coefficients for TWW (r = .48; Δr = .15, z = 2.32, p = .082), WSC (r = .54; Δr = .09, z = 1.74, p = .325), CWS (Δr = –.03, z = 0.77, p = .999), or CIWS (Δr = –.04, z = 0.49, p = .999). At spring, all coefficients were at or near the r = .60 standard, with the aLPA coefficient (r = .72) again not statistically different than the coefficients for TWW (r = .59; Δr = .13, z = 2.48, p = .053), WSC (r = .65; Δr = .07, z = 1.59, p = .448), CWS (r = .68; Δr = .04, z = 0.71, p = .999), or CIWS (r = .67; Δr = .05, z = 0.46, p = .999).

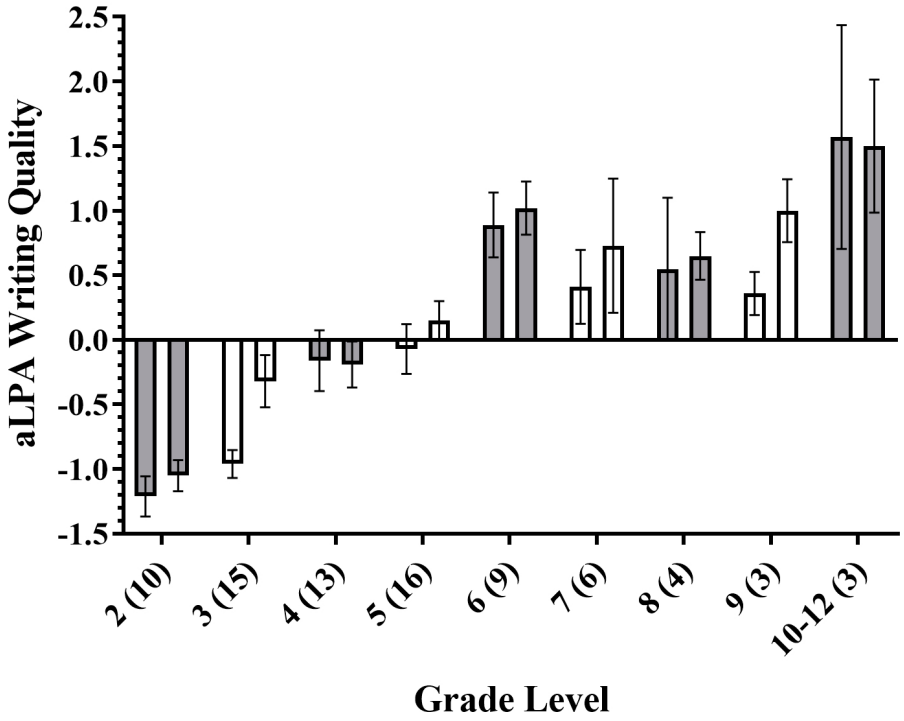
3.2.2 Story Composition

For the fall time point, the coefficients for aLPA (r = .44) and all WE-CBM scores were below the minimally sufficient standard of r = .50. There were no significant differences between the validity coefficients for aLPA and TWW (r = .34; Δr = .10, z = 1.34, p = .724), WSC (r = .36; Δr = .08, z = 1.40, p = .642), CWS (r = .45; Δr = –.01, z = 0.14, p = .999), or CIWS (r = .43; Δr = .01, z = 0.11, p = .999). For the spring time point, only the coefficient for aLPA (r = .53) was above the minimally sufficient standard; however, the aLPA coefficient again did not statistically differ from the coefficients for TWW (r = .44; Δr = .09, z = 1.31, p = .762), WSC (r = .46; Δr = .07, z = 1.22, p = .887), CWS (r = .45; Δr = .08, z = 1.13, p = .999), and CIWS (r = .38; Δr = .15, z = 1.10, p = .999).

3.3 Sensitivity to Skill Growth

For the 79 students who completed both the fall and spring narrative writing samples, aLPA scores increased from fall (M = –0.14) to spring (M = 0.11), $t(78)$ = 3.37, p = .001. Considering effect size (g = 0.25), the increase was of small magnitude. Fall and spring aLPA scores by grade level are presented in Figure 4.

Figure 4: Change in aLPA Writing Quality From Fall to Spring by Grade



Note. n = 79. aLPA scores are standardized across the fall and spring time points (M = 0, SD = 1). For each grade level, the left bar is the mean for the fall time point and the right bar is the mean for the spring time point. Values in parentheses are within-grade sample sizes.

4. Discussion

The purpose of this study was to evaluate the evidence in support of several key inferences for the use of aLPA to progress monitor the writing skills of students with parent-identified learning difficulties. First, in support of the scoring inference assumption that aLPA scores represent writing quality, we found aLPA scores to be highly correlated with quality ratings on the scored writing samples (r = .88 and .93). Although WE-CBM metrics were also highly correlated with ratings of writing quality (rs = .79 to .89), it is important to note that the writing sample duration in this study (10 minutes) was longer than in typical WE-CBM procedures (3 minutes). Second, we found partial support for the extrapolation inference assumption that aLPA scores can serve as general indicators of writing skill. In support of this assumption, aLPA scores were highly correlated, above the r = .60 criterion, with scores at both time points on a standardized writing subtest that primarily assesses spelling and grammar; in addition, the aLPA score at the spring time point was above the minimally acceptable threshold of r = .50 for the standardized sub-

test assessing substantive writing quality. There was minimal evidence of improved extrapolation for aLPA relative to WE-CBM, in large part because aLPA and WE-CBM scores were highly correlated; validity coefficients were generally very similar for aLPA and the more complex WE-CBM scores of CWS and CIWS. Third, in support of the decision inference assumption that aLPA scores are sensitive to writing skill growth for students with learning difficulties, we found small, statistically significant improvements in scores across the fall and spring of 1 academic year.

4.1 Validation of Formative Writing Assessments

Although early stage, the current study has implications for future research investigating the validity of formative writing assessments. Consistent with contemporary validation approaches (American Educational Research Association et al., 2014; Kane, 2013), we provide an interpretation and use validity argument to frame the current study and guide our future research on aLPA. In doing so, we contrast aLPA with WE-CBM; this contrast is most notable in reference to our scoring inference that scores should represent quality on the scored writing samples. Related to the behavioral orientation of foundational CBM research, CBM scores are typically assumed to be direct measures of observable behaviors (Christ et al., 2016). Although CBM scores are clearly intended to serve as general indicators of skill, which we also specify in our extrapolation inference for aLPA, there is some theoretical debate as to whether CBM scores assess the construct of fluency or indeed any latent construct (Espin & Deno, 2016). This limited attention to scoring inferences in CBM research can be problematic when WE-CBM scores are criticized as solely representing text length instead of writing quality (Gansle et al., 2002; Ritchey & Coker, 2013). Our current findings provide some evidence that WE-CBM scores assess quality on the scored writing samples, but with the caveat that we used a longer writing duration that may increase the likelihood that students had adequate time to engage in the writing process. We also provide evidence that aLPA scores are more highly correlated with quality ratings on the scored writing samples than counts of the number of total words written, thereby demonstrating that aLPA scores are assessing quality, not just composition length.

Formally specifying and evaluating the scoring inference also provides key information to interpret our extrapolation findings. Although extrapolation of aLPA to the standardized subtest assessing writing mechanics was good, extrapolation of aLPA scores to the subtest assessing substantive writing quality was not consistently adequate. Given the support for our scoring inference that aLPA scores can represent quality on the evaluated writing samples, the limited extrapolation to the substantive quality subtest appears to be an issue of limited generalization rather than problems in the aLPA scoring model – specifically, we may need to calculate aLPA scores across more than one narrative writing sample to have adequate generalization, and in turn improved extrapolation to other standardized writing assessments. We did not evaluate the generalization inference in the current study,

but other research indicates that more than one writing sample is needed for adequate generalization and reliable scores (Keller-Margulis et al., 2016; Kim et al., 2017). Also, in the current study, both the aLPA writing samples and standardized writing assessment were based on narrative compositions; evaluating generalization and extrapolation across writing genres will be necessary in future aLPA research.

4.2 Limitations

Several limitations in the current study should be considered. Because we analyzed extant data collected by the nonprofit organization's tutors as part of typical service delivery, we do not have information to support the fidelity of assessment administration procedures or have detailed information on the tutoring provided to students during the academic year. The tutors, however, did follow written assessment instructions, and we present evidence of scoring reliability for the hand scoring (WE-CBM) and quality ratings completed by the research team. In addition, although the narrative writing task used for aLPA scoring was designed by the nonprofit agency, it was a standalone assessment that was not fully embedded in their curriculum. In the future, it may be helpful to use aLPA on initial drafts from writing tasks that are later revised based on feedback and included in writing portfolios that are shared with others; by doing so, the writing task may have more value for the students, increasing motivation and eliciting better writing (Calfee & Miller, 2007). There are also some limitations related to sampling: (a) detailed demographic information for participants was not available from the nonprofit organization, which may complicate evaluations of potential generalizability to other contexts; (b) the standardized writing assessment was completed by a non-random 31% of students, contributing to wide confidence intervals around the validity coefficients that limited our ability to test differences between aLPA and WE-CBM scores; and (c) the sample size at each grade level was too small to permit within-grade level analyses, potentially leading to larger magnitude validity coefficients given that across-grade WE-CBM validity coefficients tend to be higher than within-grade coefficients (McMaster & Espin, 2007).

4.3 Conclusion

Overall, the current findings provide preliminary evidence that aLPA can be used to score narrative writing samples from students with parent-identified learning difficulties for writing quality, with preliminary evidence that aLPA scores are sensitive to student growth and some mixed evidence that aLPA scores can be extrapolated as indicators of general writing skill. Considering that correlations of aLPA scores were comparable to the best-performing hand-scored WE-CBM metrics, the use of automated text evaluation in aLPA may increase the likelihood that teach-

ers will use aLPA to monitor the written expression progress of their students due to the reduced scoring time required. Because the writeAlizer scoring model used in the current study was originally trained on writing samples from students mostly without learning difficulties, future research should investigate if training an additional scoring model on writing samples from students with learning difficulties would improve performance. In addition, future research should determine the number and duration of writing samples necessary for adequate generalization of aLPA scores, potentially improving the ability of aLPA scores to extrapolate as general indicators of writing skill. By conducting such studies in collaboration with community schools and agencies, the likelihood that aLPA can be fully embedded in the curriculum will increase, thereby making assessment more meaningful to the students and teachers.

Acknowledgments

The research reported here was supported in part by grants from the Social Sciences and Humanities Research Council of Canada and the Institute of Education Sciences, U.S. Department of Education, through Grant R305A190100. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

We would like to thank Norlan Cabot, Kate Raven, and the Learning Disabilities Association of Greater Vancouver for the research partnership that supported the current study. We would also like to thank Michèle P. Cheng, Eun Young Kwon, and Ioanna K. Tsiriotakis for their assistance in scoring writing samples and thank Daniela P. Blettner for the German translation of the abstract.

References

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

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

Calfee, R. C., & Miller, R. G. (2007). Best practices in writing assessment. In S. Graham, C. A. MacArthur, & J. Fitzgerald (Eds.), *Best practices in writing instruction* (pp. 265–286). Guilford.

Christ, T. J., van Norman, E. R., & Nelson, P. M. (2016). Foundations of fluency-based assessments in behavioral and psychometric paradigms. In K. D. Cummings & Y. Petscher (Eds.), *The fluency construct: Curriculum-based measurement concepts and applications* (pp. 143–163). Springer.

Clark, A. P., Howard, K. L., Woods, A. T., Penton-Voak, I. S., & Neumann, C. (2018). Why rate when you could compare? Using the "EloChoice" package to assess

pairwise comparisons of perceived physical strength. *PLoS ONE, 13*(1), Article e0190393. https://doi.org/10.1371/journal.pone.0190393

Dascalu, M., Crossley, S. A., McNamara, D. S., Dessus, P., & Trausan-Matu, S. (2018). Please ReaderBench this text: A multi-dimensional textual complexity assessment framework. In S. D. Craig (Ed.), *Tutoring and intelligent tutoring systems* (pp. 251–271). Nova Science.

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

Diedenhofen, B., & Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PLoS ONE, 10*(4), Article e0121945. https://doi.org/10.1371/journal.pone.0121945

Espin, C. A., & Deno, S. L. (2016). Conclusion: Oral reading fluency or reading aloud from text: An analysis through a unified view of construct validity. In K. D. Cummings & Y. Petscher (Eds.), *The fluency construct: Curriculum-based measurement concepts and applications* (pp. 365–384). Springer.

Espin, C. A., Scierka, B. J., Skare, S., & Halverson, N. (1999). Criterion-related validity of curriculum-based measures in writing for secondary school students. *Reading & Writing Quarterly: Overcoming Learning Difficulties, 15*(1), 5–27. https://doi.org/10.1080/105735699278279

Fewster, S., & Macmillan, P. D. (2002). School-based evidence for the validity of curriculum-based measurement of reading and writing. *Remedial and Special Education, 23*(3), 149–156. https://doi.org/10.1177/07419325020230030301

Filderman, M. J., Toste, J. R., Didion, L. A., Peng, P., & Clemens, N. H. (2018). Data-based decision making in reading interventions: A synthesis and meta-analysis of the effects for struggling readers. *The Journal of Special Education, 52*(3), 174–187. https://doi.org/10.1177/0022466918790001

Förster, N., & Souvignier, E. (2014). Learning progress assessment and goal setting: Effects on reading achievement, reading motivation and reading self-concept. *Learning and Instruction, 32*, 91–100. https://doi.org/10.1016/j.learninstruc.2014.02.002

Förster, N., & Souvignier, E. (2015). Effects of providing teachers with information about their students' reading progress. *School Psychology Review, 44*(1), 60–75. https://doi.org/10.17105/SPR44-1.60-75

Fuchs, D., McMaster, K. L., Fuchs, L. S., & Al Otaiba, S. (2013). Data-based individualization as a means of providing intensive instruction to students with serious learning disorders. In H. L. Swanson, K. R. Harris, & S. Graham (Eds.), *Handbook of learning disabilities* (2nd ed., pp. 526–544). Guilford.

Furey, W. M., Marcotte, A. M., Hintze, J. M., & Shackett, C. M. (2016). Concurrent validity and classification accuracy of curriculum-based measurement for written expression. *School Psychology Quarterly, 31*(3), 369–382. https://doi.org/10.1037/spq0000138

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. https://doi.org/10.1080/02796015.2002.12086169

Hammill, D. D., & Larsen, S. C. (2009). *Test of written language 4 (TOWL-4)*. Pro-Ed Assessments.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

Hosp, M. K., Hosp, J. L., & Howell, K. W. (2016). *The ABCs of CBM: A practical guide to curriculum-based measurement* (2nd ed.). Guilford.

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

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Keller-Margulis, M. A., Mercer, S. H., & Matta, M. (2021). Validity of automated text evaluation tools for written-expression curriculum-based measurement: A comparison study. *Reading and Writing: An Interdisciplinary Journal, 34*(10), 2461–2480. https://doi.org/10.1007/s11145-021-10153-6

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

Kim, Y. G., Gatlin, B., Al Otaiba, S., & Wanzek, J. (2018). Theorization and an empirical investigation of the component-based and developmental text writing fluency construct. *Journal of Learning Disabilities, 51*(4), 320–335. https://doi.org/10.1177/0022219417712016

Kim, Y. G., Schatschneider, C., Wanzek, J., Gatlin, B., & Al Otaiba, S. (2017). Writing evaluation: Rater and task effects on the reliability of writing scores for children in Grades 3 and 4. *Reading and Writing, 30*(6), 1287–1310. https://doi.org/10.1007/s11145-017-9724-6

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

McMaster, K. L., & Campbell, H. (2008). New and existing curriculum-based writing measures: Technical features within and across grades. *School Psychology Review, 37*(4), 550–556. https://doi.org/10.1080/02796015.2008.12087867

McMaster, K. L., Fuchs, D., Fuchs, L. S., & Compton, D. L. (2005). Responding to nonresponders: An experimental field trial of identification and intervention methods. *Exceptional Children, 71*(4), 445–463. https://doi.org/10.1177/001440290507100404

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

McMaster, K. L., Shin, J., Espin, C. A., Jung, P.-G., Wayman, M. M., & Deno, S. L. (2017). Monitoring elementary students' writing progress using curriculum-based measures: Grade and gender differences. *Reading and Writing, 30*(9), 2069–2091. https://doi.org/10.1007/s11145-017-9766-9

Meng, X.-I., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin, 111*(1), 172–175. https://doi.org/10.1037/0033-2909.111.1.172

Mercer, S. H. (2020). *writeAlizer: Generate predicted writing quality and written expression CBM scores* (Version 1.2.0). Retrieved from https://github.com/shmercer/writeAlizer/

Mercer, S. H., Keller-Margulis, M. A., Faith, E. L., Reid, E. K., & Ochs, S. (2019). The potential for automated text evaluation to improve the technical adequacy of written expression curriculum-based measurement. *Learning Disability Quarterly, 42*(2), 117–128. https://doi.org/10.1177/0731948718803296

Mercer, S. H., Martínez, R. S., Faust, D., & Mitchell, R. R. (2012). Criterion-related validity of curriculum-based measurement in writing with narrative and expository prompts relative to passage copying speed in 10th grade students. *School Psychology Quarterly, 27*(2), 85–95. https://doi.org/10.1037/a0029123

Morgan, P. L., & Sideridis, G. D. (2006). Contrasting the effectiveness of fluency interventions for students with or at risk for learning disabilities: A multilevel random coefficient modeling meta-analysis. *Learning Disabilities Research & Practice, 21*(4), 191–210. https://doi.org/10.1111/j.1540-5826.2006.00218.x

National Center on Intensive Intervention. (2018). *Academic screening tools chart rating rubric*. Retrieved from https://intensiveintervention.org/sites/default/files/NCII_AcademicScreening_RatingRubric_July2018.pdf

Neumann, C. (2019). *EloChoice: Preference rating for visual stimuli based on Elo ratings* (Version 0.29.4). Retrieved from https://CRAN.R-project.org/package=EloChoice

Pelánek, R. (2016). Applications of the Elo rating system in adaptive educational systems. *Computers & Education, 98*, 169–179. https://doi.org/10.1016/j.compedu.2016.03.017

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

Ritchey, K. D., & Coker, D. L., Jr. (2013). An investigation of the validity and utility of two curriculum-based measurement writing tasks. *Reading & Writing Quarterly: Overcoming Learning Difficulties, 29*(1), 89–119. https://doi.org/10.1080/10573569.2013.741957

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

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

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*(4), 273–286. https://doi.org/10.1037/h0070288

Videen, J., Deno, S. L., & Martson, D. (1982). *Correct word sequences: A valid indicator of proficiency in written expression*. University of Minnesota, Institute for Research on Learning Disabilities.

Werb, J. (2007, October 31). *Provincial government fails learning-disabled kids*. The Georgia Straight. https://www.straight.com/article-116366/provincial-government-fails-learning-disabled-kids

Wilson, J. (2018). Universal screening with automated essay scoring: Evaluating classification accuracy in grades 3 and 4. *Journal of School Psychology, 68*, 19–37. https://doi.org/10.1016/j.jsp.2017.12.005

Wilson, J., Chen, D., Sandbank, M. P., & Hebert, M. (2019). Generalizability of automated scores of writing quality in Grades 3–5. *Journal of Educational Psychology, 111*(4), 619–640. https://doi.org/10.1037/edu0000311