https://doi.org/10.1177/0829573520987753

Accuracy of Automated Written Expression Curriculum-based Measurement Scoring

Sterett H. Mercer[1], Joanna E. Cannon[1], Bonita Squires[1,2], Yue Guo[1], and Ella Pinco[1]

[1]Faculty of Education, University of British Columbia

[2]Faculty of Health, Dalhousie University

Author Note

Sterett H. Mercer https://orcid.org/0000-0002-7940-4221

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 Eun Young Kwon, and Ioanna K. Tsiriotakis for their assistance in scoring writing samples

Correspondence concerning thus article should be addressed to Sterett H. Mercer, 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

Abstract

We examined the extent to which automated written expression curriculum-based measurement (aWE-CBM) can be accurately used to computer score student writing samples for screening and progress monitoring. Students (*n* = 174) with learning difficulties in Grades 1–12 who received 1:1 academic tutoring through a community-based organization completed narrative writing samples in the fall and spring across two academic years. The samples were evaluated using four automated and hand-calculated WE-CBM scoring metrics. Results indicated automated and hand-calculated scores were highly correlated at all four timepoints for counts of total words written (*r*s = 1.00), words spelled correctly (*r*s = .99 – 1.00), correct word sequences (CWS; *r*s = .96 – .97), and correct minus incorrect word sequences (CIWS; *r*s = .86 – .92). For CWS and CIWS, however, automated scores systematically overestimated hand-calculated scores, with an unacceptable amount of error for CIWS for some types of decisions. These findings provide preliminary evidence that aWE-CBM can be used to efficiently score narrative writing samples, potentially improving the feasibility of implementing multi-tiered systems of support in which the written expression skills of large numbers of students are screened and monitored.

Keywords: written expression, curriculum-based measurement, automated text evaluation, screening, progress monitoring

Accuracy of Automated Written Expression Curriculum-based Measurement Scoring

To effectively support students with learning difficulties, measures that can identify which students need additional assistance and that can progress monitor the effectiveness of academic interventions are needed (Jung et al., 2018). Since the 1970s, researchers in special education have examined curriculum-based measurement (CBM) for these purposes. CBMs are brief and efficient assessments that can be administered frequently during instruction, with evidence of reliability and validity for defensible decisions about student progress during instructional interventions (Deno, 1985). For example, brief assessments of oral passage reading (i.e., number of words read correctly in 1 minute) work well as a technically adequate indicator of overall reading proficiency (Reschly et al., 2009), are sensitive to reading skill growth during intervention (Morgan & Sideridis, 2006), and improve student outcomes when used by teachers during intervention (Filderman et al., 2018). Compared to reading CBM, research on written expression CBM (WE-CBM) is less well developed (Tindal, 2013) despite the importance of writing skills for students' academic and occupational success (National Commission on Writing, 2004).

Administration of WE-CBM typically includes presentation of a short story starter (e.g., "One day on the way to school, I..."), and then student generation of a three- to five- minute writing sample following a one-minute planning time (Hosp et al., 2016). Multiple WE-CBM metrics are utilized to score samples, for example, the total number of words written (TWW), counts of words spelled correctly (WSC), counts of correct words sequences (CWS, the number of adjacent words that are syntactically and semantically acceptable in context and spelled and punctuated correctly; Videen et al., 1982), and counts of correct minus incorrect word sequences (CIWS; Espin et al., 2000). Although U.S. based norms and data management are available for

TWW, WSC, and CWS through the aimsweb platform (http://www.aimsweb.com), to our knowledge, no comparable norms or data management platforms are available in Canada, and no written expression assessments are listed as meeting the National Center on Intensive Intervention's (http://intensiveintervention.org/) standards for reliability and validity of screening and progress monitoring tools.

Two key challenges have hindered the development and use of WE-CBM beyond the early elementary grades (for a review of CBM for beginning writers, see Ritchey et al., 2016). First, studies have found, using generalizability theory, that multiple, longer-duration writing samples are needed to obtain adequate reliability for WE-CBM in Grades 2–5 (Keller-Margulis et al., 2016; Kim et al., 2017). Second, a recent meta-analysis found that more complex WE-CBM scores (CWS and CIWS) had higher validity coefficients than simpler WE-CBM metrics (TWW and WSC) at all grade levels from K to 12 (Romig et al., 2017). In combination, the need for longer, multiple writing samples and more complex scoring approaches can limit the feasibility of WE-CBM (Espin et al., 1999), particularly for screening when writing samples are obtained from all students in multiple grades simultaneously.

To address these feasibility concerns, several studies have investigated automated text evaluation to score writing samples used for universal screening (Mercer et al., 2019; Wilson, 2018). In Wilson (2018), the commercial Project Essay Grade program (PEG; Page, 2003) was used to score 60–minute argumentative writing samples from students in Grades 3 and 4. PEG Total scores, formed from the sum of five-point analytic rubric ratings on six writing dimensions (development of ideas, organization, style, sentence structure, conventions, and word choice) had good diagnostic accuracy in predicting whether students met proficiency standards on a state-mandated English Language Arts assessment. Similarly, Mercer et al. (2019) investigated the extent to which composite scoring models based on Coh-Metrix (Graesser et al., 2014), a free program originally designed to predict text readability, could predict holistic writing quality on seven-minute screening samples from students in Grades 2–5. Results were that both the composites based on Coh-Metrix scores and typical WE-CBM scores correlated with holistic writing quality at r = .73 – .77. In both studies, however, automated text evaluation was used to generate holistic writing quality scores, either as the sum of scores across analytic rating dimensions (Wilson, 2018) or through a composite scoring model (Mercer et al., 2019), rather than to directly generate WE-CBM scores. By simplifying the scoring process, using automated text evaluation for WE-CBM (aWE-CBM) could potentially remove feasibility barriers to the use of WE-CBM in schools for data-based screening and progress monitoring decisions.

Current Study

The purpose of the current study is to determine the accuracy of automated scoring for the most frequently used WE-CBM metrics: TWW, WSC, CWS, and CIWS. Specifically, we address the following research questions with narrative writing samples from Grade 1–12 students with learning difficulties:

1. How strongly do aWE-CBM scores relate to hand-calculated WE-CBM scores? To address this question, we calculate Pearson r correlations, with expectations that correlations would be nearly perfect, $r \geq .90$.
2. How precise are aWE-CBM scores? To address this question, we examine root mean square error (*RMSE*) values, with values closer to zero indicating smaller discrepancies between hand-calculated and automated scores. In addition to being useful in practice to form confidence intervals around predicted WE-CBM scores (± 2**RMSE* is a simple approximation of a 95% prediction interval), *RMSE* values can be divided by the standard

deviation of test scores as an indicator of reliability; a standard error of less than one-third the magnitude of the standard deviation of test scores has been recommended as a minimal standard for important applied decisions (Nunnally, 1978).

3. To what extent are aWE-CBM scores unbiased? To address this question, we conduct paired sample *t* tests to investigate the extent to which aWE-CBM systematically over- or underestimates WE-CBM scores.

Method

Participants and Setting

For approximately two hours per week, all participants received 1:1 tutoring by a community-based non-profit organization. In order to track progress and inform further instruction, the organization collected picture-prompted narrative writing samples at two time points per year, the fall (September–October) and spring (April–May). For the 2017–2018 school year, 106 student participants in Grades 2–12 completed at least one writing sample and 40% were female. The fall writing sample was completed by 103 of the participants and the spring sample by 83 participants. The majority of participants attended elementary school (Grades 2–7; *n* = 85) and the remainder (*n* = 21) attended secondary school (Grades 8–12). For the 2018–2019 school year, 68 students in Grades 1–12 participated, with 51 and 52 students completing the fall and spring writing samples, respectively. Eighty–four percent of students were in the elementary grades (1–7), and 41% were female.

Detailed demographic and disability information about the participants is not provided because we only had access to extant writing samples. Common demographics among all participants included the fact that their parents sought community-based tutoring to provide them with support beyond what their school offered. Most participants attended an urban, culturally and linguistically diverse school district in Canada with approximately 52,000 students, 44% of which reported speaking a language other than English at home. Of the 160 different home languages within the district, the top five include: Cantonese (17%), Mandarin (11%), Tagalog (5%), Vietnamese (4%), and Punjabi (4%). Some students in the district, approximately 17%, were eligible for English language supports, and another 11% received special education services.

Measures

Writing samples were collected by the organization's tutors by presenting participants with an array of travel, recreation, and lifestyle magazine photos (e.g., amusement park rides, animals, restaurants). The participants selected one picture as a writing prompt, and were allowed 10 minutes to handwrite a composition with no help from the tutors. Consistent with Behavioural Research Ethics Board approved procedures, students' parents or guardians were asked by the organization for consent to release de-identified writing samples to the research team. Before scoring, all samples were typed, preserving errors in spelling and grammar, by a member of the research team, with the accuracy of all transcriptions verified by another member of the team. Following transcription, we scored these narrative writing samples for hand-calculated and automated WE-CBM metrics.

Hand-calculated WE-CBM

Based on the Hosp et al. (2016) guidelines, four hand-calculated WE-CBM metrics (TWW, WSC, CWS, CIWS) were scored. For TWW, we counted the total number of one or more letters that were separated by spaces, even if these words were used mistakenly in context or misspelled. To calculate WSC, we counted correctly spelled English words regardless of the context. For CWS, we counted each sequence of two adjacent words that were spelled correctly and were syntactically and semantically acceptable in context; correct punctuation and capitalization were also considered. For CIWS, we calculated the difference between correct and incorrect word sequences. Forty-two percent of the writing samples were independently scored by two raters. Agreement was very strong between the raters for all metrics: TWW (*r* = 1.00), WSC (*r* = 1.00), CWS (*r* = 1.00), and CIWS (*r* = .99).

Automated WE-CBM

We used the open-source writeAlizer R package (Mercer, 2020) to generate aWE-CBM scores based on the output of a text analysis program, Grammar And Mechanics Error Tool (GAMET; Crossley et al., 2019). GAMET is a free program, based on the open-source LanguageTool application (https://languagetool.org/), that batch processes text files to generate the following metrics: (a) word count, (b) misspellings, (c) grammatical errors, (d) duplication errors (e.g., “I made made an error.”), (e) typography errors including capitalization and punctuation, and (f) white space errors such as inappropriate spacing before punctuation or between words. In addition, writeAlizer generates percentages of misspelled words and grammatical errors by dividing these counts by the total word count. For automated TWW scores, writeAlizer uses the word count score generated by GAMET. For automated WSC, writeAlizer subtracts GAMET-identified misspellings from the word count. Automated CWS and CIWS scores are based on ensembles of four machine learning algorithms that were trained on 7 min narrative writing samples from students in Grades 2–5 (see Mercer et al., 2019, for sample description); the weightings of each of the GAMET metrics, overall and in the individual algorithms, plus the weightings of each algorithm in the CWS and CIWS scoring models are presented in Table 1. More details on the algorithms listed in Table 1 are available in Hastie et al. (2009).

Results

Means and standard deviations for all aWE-CBM and WE-CBM scores by timepoint are presented in Table 2. Below, we present results by WE-CBM metric. Complete results, including 95% confidence intervals for correlations and full details of the paired sample *t* tests, are presented in Table 3.

Total Words Written

At all four timepoints, automated TWW scores were perfectly correlated with hand-calculated TWW scores at $r = 1.00$. *RMSE* values (.47 – 2.98) were small relative to mean TWW scores (58.03 – 79.10) and also well below Nunnally's (1978) recommendation that error magnitude be less than one-third of the TWW score standard deviations, with proportions of .07, .04, .01, and .07, respectively, by timepoint. Automated TWW scores were not statistically different from hand-calculated TWW scores at any timepoint, indicating no systematic over- or under-estimation of scores.

Words Spelled Correctly

Automated WSC scores were nearly perfectly correlated ($r = .99 – 1.00$) with hand-calculated WSC scores at all timepoints. *RMSE* values (1.69 – 4.77) were small relative to mean WSC scores (50.49 – 71.75) and also small relative to the WSC standard deviations, with proportions of .12, .08, .05, and .09, respectively, by timepoint. Automated WSC scores were not statistically different from hand-calculated WSC scores for three of the four timepoints. In fall of 2018–2019, automated scores were on average .98 below hand-calculated WSC ($p < .001$), but this difference was of trivial magnitude ($d = .03$).

Correct Word Sequences

Correlations between automated and hand-calculated CWS scores were nearly perfect at all timepoints (*r* = .96 – .97). *RMSE* values (9.16 – 12.15) were moderate relative to the CWS means (38.37 – 59.79) but less than one-third of the CWS standard deviations at all timepoints, with proportions of .26, .23, .29, and .29, respectively. There was a tendency toward over-estimating hand-calculated CWS; automated CWS scores were significantly higher at all timepoints (*p* < .05) by 2.30 to 6.65 CWS. These differences were of small magnitude with *d* = .18, .08, .27, and .09, respectively, by timepoint.

Correct Minus Incorrect Word Sequences

Correlations between automated and hand-calculated CIWS scores were very strong (*r* = .86 – .92), but noticeably lower than for other WE-CBM metrics. *RMSE* values (16.77 – 20.48) were moderate relative to the adjusted CIWS means (68.12 – 120.81) and above the one-third recommendation for the magnitude of error relative to the CIWS standard deviation, with proportions of .51, .40, .44, and .46 by timepoint. Automated CIWS scores were significantly greater than hand-calculated CIWS at all timepoints (*p* < .01) by 9.55 to 16.27 CIWS. These differences were of small to moderate magnitude with *d* = .41, .20, .43, and .22, respectively, by timepoint.

Discussion

The purpose of this study was to evaluate the accuracy of automated text evaluation as an alternative to hand-calculated WE-CBM scores. Overall, accuracy differed by WE-CBM metric. For TWW and WSC, automated and hand-calculated scores were nearly perfectly correlated at all timepoints (*r*s = .99 – 1.00), with very little prediction error and little evidence of systematic over- or underestimation of scores. Similarly, automated CWS scores were nearly perfectly correlated with hand-calculated CWS scores (*r*s = .96 – .97), but with some small overestimation of scores. By contrast, automated CIWS scores were less strongly correlated with hand-calculated scores (*r*s = .86 – .92), with evidence of small to moderate overestimation of scores contributing to *RMSE* values of approximately half the size of the CIWS score standard deviations.

Utility of aWE-CBM for Relative and Absolute Decisions

To interpret these accuracy statistics for aWE-CBM, it is helpful to differentiate between relative and absolute decisions (see Shavelson & Webb, 1991). In relative decisions, the primary purpose is to rank order students, with the specific scaling of scores not important. In educational contexts, screening decisions are typically relative, specifically when a certain percentage of students are identified as having *at risk* status based on their performance relative to peers. Research with correlational designs would also involve relative decisions because variable scaling would not be important. The correlations between automated and hand-calculated scores provide strong to very strong evidence that all aWE-CBM scores can be used as a substitute for hand-calculated WE-CBM scores for relative decisions. In absolute decisions, the scaling of scores is important—some examples of absolute decisions are comparing student score growth over time, comparing average performance between groups, or comparing student scores to specific benchmark scores. In the current study, the *RMSE* values and *t* tests inform the appropriateness of aWE-CBM scores for absolute decisions. In general, there is strong evidence for automated TWW and WSC scores as substitutes for hand-calculated scores in absolute decisions and also good evidence that automated CWS scores can be used for absolute decisions, with the caveat that CWS scores may be somewhat overestimated. By contrast, the greater overestimation of CIWS scores plus the high ratio of *RMSE* to *SD* indicate that automated CIWS scores should not be used for absolute decisions.

Limitations

Limitations of the current study include (a) limited information on fidelity of assessment administration procedures, (b) unavailability of detailed demographic data, and (c) the small numbers of students at each grade level. Fidelity of administration procedures was not documented—the tutors who administered the writing samples followed the non-profit organization's written assessment instructions; however, we only had access to the writing samples the organization provided. Although this is a limitation for internal validity, it may increase external validity because educators are likely to administer writing prompts in varied ways across different settings. Another limitation is that detailed demographic information about the participants was unavailable, including whether students had a formal disability designation and to what extent students received school-based special education services. A final limitation included the small numbers of students at each grade level, precluding separate analyses by grade level. Considering that WE-CBM validity coefficients tend to be smaller when based on within-grade compared to across-grade analyses (McMaster & Espin, 2007), additional research will be needed to evaluate validity of aWE-CBM scoring at specific grade levels.

Future Research

The current findings highlight several areas to be addressed in future research. First, our finding that automated CWS and CIWS scoring tended to systematically overestimate hand-calculated values may be related to the use of output from the GAMET program for the aWE-CBM scoring models. A prior study (Crossley et al., 2019) comparing GAMET and human scoring of grammatical and mechanical errors on essays from secondary students and adults found that although the majority of errors found by GAMET were rated as accurate and meaningful, GAMET failed to identify a large number of errors identified by human raters, particularly concerning punctuation. For this reason, more research to refine the accuracy of the underlying text analysis applications (GAMET and LanguageTool) will be needed to improve aWE-CBM scoring of CWS and CIWS.

Second, although accuracy of aWE-CBM scoring for TWW, WSC, and CWS was good in the current study, it is notable that the writeAlizer scoring models were trained based on shorter-duration writing samples from Grade 2–5 students largely without disabilities (only 6% of students in Mercer et al., 2019), in contrast to the current sample of Grade 1–12 students with substantial learning difficulties. It is possible that the accuracy of aWE-CBM varies depending on the writing skill level of students, and differences in writing skill levels between the training and current samples may have contributed to the overestimation of CWS and CIWS scores. Future efforts to train aWE-CBM scoring models on samples from a more diverse set of learners may further improve scoring accuracy.

Third, our research thus far has focused on the accuracy of the scoring model, and additional software development will be necessary before aWE-CBM is ready for more widespread use. In the current study, we transcribed handwritten samples, batch submitted the samples for processing in GAMET, and then imported the GAMET output into the R program to generate predicted WE-CBM scores using syntax. These steps are likely to be too complex for the average potential user. As we continue our work in this area, we plan to (a) continue to refine and evaluate the scoring models, (b) develop software to simply the workflow of generating scores from samples, (c) explore options for online data management and report generation, and (d) establish norms and standards for performance. All of these steps will need to be completed for aWE-CBM to be practically and feasibly used for screening decisions in schools.

Fourth, and most importantly, additional research is needed on the validity of WE-CBM, whether scored by hand or computer. In a meta-analysis of criterion-related validity evidence for WE-CBM (Romig et al., 2017), only the CIWS metric approached the $r \geq .60$ validity standard for screening tools of the National Center for Intensive Intervention. Because most of the studies included in the meta-analysis did not base student skill estimates on multiple, longer duration samples, which have been identified as necessary for adequate WE-CBM reliability (Keller-Margulis et al., 2016; Kim et al., 2017), these validity coefficients are difficult to interpret. For this reason, future research will need to be conducted to investigate the validity of WE-CBM when based on more substantial samples of student writing, and it is possible that aWE-CBM scoring may improve scoring feasibility in these research efforts.

Relevance to the Practice of School Psychology

The present study demonstrates that automated scoring of TWW, WSC, and CWS can be accurately used in place of WE-CBM hand scoring, and these efficiency gains may support more widespread implementation of response-to-intervention (RTI) models and multi-tiered systems of support (MTSS; Jimerson et al., 2016) to support student writing skills. In these models, all students should be regularly screened to identify students in need of more support, with data used to determine if provided supports are adequately supporting student skill development (Jung et al., 2018). Although screening assessments in written expression are relatively brief to administer, the time required to hand score WE-CBM metrics can be substantial. In the current study, we did not collect systematic data on hand or automated scoring time for each 10-minute sample; however, prior studies have reported average WE-CBM hand scoring times of 2.0–2.5 minutes per sample for 3-minute samples (Gansle et al., 2002; Malecki & Jewell, 2003). Considering that multiple, longer-duration writing samples per student are needed for reliable estimates of student writing skill (Keller-Margulis et al., 2016; Kim et al., 2017), the time to score writing samples from students in multiple classes and grades, as is done in universal screening, is likely to a barrier to more widespread adoption of these practices (Espin et al, 1999).

As emphasized in practice guidelines in Canada (e.g., Ontario Psychological Association Section on Psychology in Education, 2013) and the United States (National Association of School Psychologists, 2017), school psychologists have the data-based decision making and instructional consultation skills to support these screening and monitoring practices as part of MTSS/RTI, and an automated scoring option may facilitate these efforts. Given that, to our knowledge, only one U.S.-based provider offers norms and data management options for WE-CBM and no Canada-based norms or providers are available, aWE-CBM scoring may also facilitate the widespread data collection needed to conduct the validity studies and develop the local norms (Patton et al., 2014) that will be required to use these scores for defensible instructional decisions.

References

Crossley, S. A., Bradfield, F., & Bustamante, A. (2019). Using human judgments to examine the validity of automated grammar, syntax, and mechanical errors in writing. *Journal of Writing Research, 11*, 251–270. https://doi.org/10.17239/jowr-2019.11.02.01

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

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

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*, 5–27. https://doi.org/10.1080/105735699278279

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*, 174–187. https://doi.org/10.1177/0022466918790001

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*, 477–497.

Graesser, A. C., McNamara, D. S., Cai, Z., Conley, M., Li, H., & Pennebaker, J. (2014). Coh-Metrix measures text characteristics at multiple levels of language and discourse. *The Elementary School Journal, 115*, 210–229. https://doi.org/10.1086/678293

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. https://doi.org/10.1007/b94608

Jimerson, S. R., Burns, M. K., & VanDerHeyden, A. M. (2016). *Handbook of response to intervention: The science and practice of multi-tiered systems of support* (2nd ed.). Springer. https://doi.org/10.1007/978-1-4899-7568-3

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*, 144–155. https://doi.org/10.1111/ldrp.12172

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

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*, 1287–1310. https://doi.org/10.1007/s11145-017-9724-6

Malecki, C. K., & Jewell, J. (2003). Developmental, gender, and practical considerations in scoring curriculum-based measurement writing probes. *Psychology in the Schools, 40*, 379–390. https://doi.org/10.1002/pits.10096

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

Mercer, S. H. (2020). *writeAlizer: Generate predicted writing quality and written expression CBM scores*. (Version 1.2.0) [Computer software]. 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*, 117–128. https://doi.org/10.1177/0731948718803296

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*, 191–210. https://doi.org/10.1111/j.1540-5826.2006.00218.x

National Association of School Psychologists. (2017). *Leveraging essential school practices, ESSA, MTSS, and the NASP practice model: A crosswalk to help every school and student succeed*. Author. http://www.nasponline.org/practice-model/ESSA-MTSS-crosswalk

National Commission on Writing. (2004). *Writing: A ticket to work... or a ticket out: A survey of business leaders.* College Board. https://archive.nwp.org/cs/public/print/resource/2540

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.

Ontario Psychological Association Section on Psychology in Education. (2013). *Professional practice guidelines for school psychologists in Ontario*. Ontario Psychological

Association. https://doi.org/http://psych.on.ca/OPA/media/Public/OPA%20Guidelines%20and%20Reviews/professional-practice-guidelines-for-school-psychologists-in-ontario-2013.pdf

Page, E. B. (2003). Project Essay Grade: PEG. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43–54). Lawrence Erlbaum Associates.

Patton, K. L. S., Reschly, A. L., & Appleton, J. (2014). Curriculum-based measurement as a predictor of performance on a state assessment: Diagnostic efficiency of local norms. *Educational Assessment, 19*, 284-301. https://doi.org/10.1080/10627197.2014.964117

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*, 427–469. https://doi.org/10.1016/j.jsp.2009.07.001

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. https://doi.org/10.1007/978-1-4939-2803-3_2

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*, 72–82. https://doi.org/10.1177/0022466916670637

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Sage.

Tindal, G. (2013). Curriculum-based measurement: A brief history of nearly everything from the 1970s to the present. *ISRN Education, 2013*, 1–29. https://doi.org/10.1155/2013/958530

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.

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

Table 1

Weightings of GAMET Metrics in writeAlizer Scoring Models by Algorithm

Metric	Overall	GBM	SVM	ENET	MARS
Correct Word Sequences Model					
Word Count	75.48	86.79	67.10	77.17	77.84
Spelling	14.26	0.62	0.00	21.41	22.05
%Spelling	8.78	12.28	27.95	0.40	0.11
Grammar	0.85	0.05	2.77	0.11	0.00
%Grammar	0.01	0.06	0.01	0.00	0.00
Duplication	0.04	0.12	0.12	0.00	0.00
Typography	0.38	0.08	1.33	0.00	0.00
White Space	0.20	0.00	0.71	0.92	0.00
Correct Minus Incorrect Word Sequences Model					
Word Count	55.60	55.76	47.57	61.43	61.35
Spelling	19.25	1.48	6.57	35.80	35.04
%Spelling	22.31	41.99	42.74	0.00	0.00
Grammar	0.82	0.00	1.69	0.00	0.62
%Grammar	0.04	0.23	0.00	0.00	0.00
Duplication	0.28	0.10	0.76	0.00	0.00
Typography	1.37	0.41	0.07	1.55	2.97
White Space	0.34	0.04	0.60	1.22	0.00

Note. The weightings sum to 100 for each model; thus, they can be viewed as the percentage contribution of each metric to the predicted scores. Overall = the ensemble model of all algorithms, GBM = stochastic gradient boosted regression trees, SVM = support vector machines (radial kernel), ENET = elastic net regression, MARS = bagged multivariate adaptive regression splines. The following regression equation was used to weight the algorithms in the CWS ensemble model: .162 + .074*GBM + .281*SVM + .001*ENET + .642*MARS. The following equation was used for the CIWS model: -.170 + .180*GBM + .346*SVM + .100*ENET + .375*MARS.

Table 2

Descriptive Statistics for Automated and Hand-calculated WE-CBM Scores by Timepoint

Metric	2017-2018				2018-2019			
	Fall (*n* = 103)		Spring (*n* = 83)		Fall (*n* = 51)		Spring (*n* = 52)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
TWW	58.03	40.07	71.93	47.20	64.65	34.13	79.10	43.12
aTWW	58.41	40.31	71.70	47.20	64.71	34.16	79.54	43.28
WSC	50.49	40.25	64.80	46.72	56.82	32.10	71.75	42.23
aWSC	50.89	39.39	64.27	46.45	55.84	31.45	71.31	41.86
CWS	38.37	36.03	53.95	47.53	41.29	31.27	59.79	42.31
aCWS	44.78	37.35	57.56	42.94	49.60	30.48	63.76	37.36
CIWS	12.77	35.59	28.81	51.16	13.12	37.89	33.44	43.54
aCIWS	27.35	35.93	39.00	41.70	29.39	31.75	42.99	36.19

Note. TWW = total words written, aTWW = automated total words written, WSC = words spelled correctly, aWSC = automated words spelled correctly, CWS = correct word sequences, aCWS = automated correct word sequences, CIWS = correct minus incorrect word sequences, aCIWS = automated correct minus incorrect word sequences.

Table 3

Correlations and Tests of Differences between Automated and Hand-calculated WE-CBM Scores

Metric	Correlations				Paired-sample *t* tests			
	r	95% CI for *r*		*RMSE*	$M_A - M_H$	*t*	*df*	*p*
				Fall 2017-2018 (*n* = 103)				
TWW	1.00	1.00	1.00	2.72	.38	1.41	102	.162
WSC	.99	.99	1.00	4.77	.41	.87	102	.389
CWS	.97	.95	.98	9.50	6.42	6.65	102	<.001
CIWS	.86	.80	.90	18.30	14.58	7.80	102	<.001
				Spring 2017-2018 (*n* = 83)				
TWW	1.00	1.00	1.00	1.71	-.23	1.23	82	.223
WSC	1.00	1.00	1.00	3.72	-.53	1.31	82	.196
CWS	.97	.96	.98	11.13	3.61	2.85	82	.006
CIWS	.92	.87	.95	20.48	10.19	4.42	82	<.001
				Fall 2018-2019 (*n* = 51)				
TWW	1.00	1.00	1.00	.47	.06	.90	50	.371
WSC	1.00	1.00	1.00	1.69	-.98	3.93	50	<.001
CWS	.96	.93	.98	9.16	8.31	6.53	50	<.001
CIWS	.90	.83	.94	16.77	16.27	6.93	50	<.001
				Spring 2018-2019 (*n* = 52)				
TWW	1.00	1.00	1.00	2.98	.44	1.08	51	.286
WSC	1.00	.99	1.00	3.80	-.44	.85	51	.402
CWS	.96	.93	.98	12.15	3.97	2.30	51	.025
CIWS	.89	.82	.94	19.81	9.55	3.48	51	.001

Note. M_A = mean of automated WE-CBM scores, M_H = mean of hand-calculated WE-CBM scores, TWW = total words written, WSC = words spelled correctly, CWS = correct word sequences, CIWS = correct minus incorrect word sequences.