This is the accepted version of the article accepted for publication in *Learning Disability Quarterly*. The final, published version of the article can be obtained on the publisher's website at https://doi.org/10.1177/0731948718803296

The Potential for Automated Text Evaluation to Improve the Technical Adequacy of Written Expression Curriculum-Based Measurement

Sterett H. Mercer

University of British Columbia

Milena A. Keller-Margulis

Erin L. Faith

Erin K. Reid

Sarah Ochs

University of Houston

Author Note

This research was supported by a College of Education Faculty Research Grant Award at the University of Houston, and the Social Sciences and Humanities Research Council of Canada.

Correspondence concerning this 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

Written-expression curriculum-based measurement (WE-CBM) is used for screening and progress monitoring students with or at risk of learning disabilities (LD) for academic supports; however, WE-CBM has limitations in technical adequacy, construct representation, and scoring feasibility as grade level increases. The purpose of this study was to examine the structural and external validity of automated text evaluation with Coh-Metrix vs. traditional WE-CBM scoring for narrative writing samples (7 min duration) collected in fall and winter from 144 second through fifth grade students. Seven algorithms were applied to train models of Coh-Metrix and traditional WE-CBM scores to predict holistic quality of the writing samples as evidence of structural validity; then, external validity was evaluated via correlations with rated quality on other writing samples. Key findings were that (a) structural validity coefficients were higher for Coh-Metrix compared to traditional WE-CBM but similar in the external validity analyses, (b) external validity coefficients were higher than reported in prior WE-CBM studies with holistic or analytic ratings as a criterion measure, and (c) there were few differences in performance across the predictive algorithms. Overall, the results highlight the potential use of automated text evaluation for WE-CBM scoring. Implications for screening and progress monitoring are discussed.

Keywords: curriculum-based measurement, writing, automated text evaluation, validity, universal screening, progress monitoring

The Potential for Automated Text Evaluation to Improve the Technical Adequacy of Written Expression Curriculum-Based Measurement

Writing is a critical skill for success in school, higher education, and the workforce (Salahu-Din, Persky, & Miller, 2007). Despite the recognized value of writing, data from the National Assessment of Educational Progress (National Center for Education Statistics, 2012) indicate that 73% of children in both eighth and twelfth grades are not proficient writers and thus are not prepared for post-secondary work. To address these troubling statistics, teachers need a tool for measuring writing skills in an accurate and efficient manner so that students with or at-risk for learning disabilities (LD) in written expression can be screened for intervention and the effectiveness of these efforts can be progress monitored. An existing tool, written-expression curriculum-based measurement (WE-CBM), can be used to meet this need; however, the validity of WE-CBM is questionable with a weak mean validity coefficient of $r = .55$ reported in a recent meta-analysis (Romig, Therrien, & Lloyd, 2016), with some evidence that validity coefficients tend to decrease as student writing becomes more complex in the upper elementary grades and beyond (McMaster & Espin, 2007).

Non-optimal validity coefficients for static WE-CBM scores, as reported in Romig et al. (2016), are problematic because they suggest that screening decisions using these data (which students are at risk for LD and need additional assistance?) may not be sufficiently defensible; also, when the technical adequacy of static CBM scores is questionable, the ability to reliably and validly assess skill growth in progress monitoring will also be hindered (Silberglitt, Parker, & Muyskens, 2016), thereby limiting the defensibility of decisions about response to instruction and intervention for students with or at risk of LD. The purpose of this study was to examine the

potential of automatic text evaluation for WE-CBM scoring to improve validity by capturing not only word-level, but also sentence- and discourse-level elements of writing.

WE-CBM Technical Adequacy

WE-CBM was developed as a simple, efficient, and repeatable assessment approach to screen and monitor the writing performance of students with or at risk of LD with an emphasis on reliability and validity so that decisions about risk status and response to instruction are defensible (Deno, 1985). Early WE-CBM studies indicated adequate reliability and validity for short duration samples (Deno, Marston, & Mirkin, 1982; Marston & Deno, 1981); however, these early findings, particularly those related to validity, have proven difficult to replicate (McMaster & Espin, 2007). Validity studies for WE-CBM largely indicate results in the weak to moderate range (McMaster & Espin, 2007; Romig et al., 2016). Studies of duration typically include 3, 5, 7, and 10 min samples of writing, and generally find that longer durations provide improved technical adequacy (e.g., Espin et al., 2008; Weissenburger & Espin, 2005), but reduce feasibility due to additional administration and scoring time (Espin, Scierka, Skare, & Halverson, 1999; Gansle, Noell, VanDerHeyden, Naquin, & Slider, 2002).

Concerns with validity have resulted in the proliferation of additional metrics that are largely variations of the original countable indices. WE-CBM metrics can be grouped as production-dependent, production-independent, and accurate-production metrics (Malecki & Jewell, 2003). Production-dependent metrics include the original WE-CBM metrics such as Total Words Written (TWW) and Correct Word Sequences (CWS), and production-independent metrics include variations of these such as percent CWS (%CWS) to control for variation in the amount of writing produced. The accurate-production metric of correct minus incorrect writing sequences (CIWS) combines accuracy and fluency (Espin et al., 2000; Espin et al., 2008), and

although validity findings tend to be higher than for other metrics (Mercer, Martínez, Faust, & Mitchell, 2012; Romig et al., 2016; Weissenburger & Espin, 2005), this metric requires significant time to reliably score and thus may be less feasible for use in universal screening (Espin et al., 1999; Gansle et al., 2002). Work in this area has largely investigated the technical adequacy of individual metrics such as TWW vs. CIWS (McMaster & Espin, 2007); however, given that these metrics, when calculated on the same samples, are moderately to highly correlated, composites based on multiple WE-CBM metrics would likely improve reliability and validity (Codding, Petscher, & Truckenmiller, 2015; Espin et al., 1999).

Construct Representation in WE-CBM

Construct representation, i.e., the extent to which administration procedures and scoring methods adequately assess important dimensions of the writing quality construct (Messick, 1995), has received limited attention in WE-CBM. Writing is a complex activity that involves numerous processes such as planning and generating ideas for text, transcribing the ideas to paper or via keyboard, and various other cognitive activities such as self-regulation (see Berninger & Amtmann, 2003, for a model of writing incorporating these elements). Current WE-CBM metrics largely capture transcription skills at the word and sentence levels of language, which may partly explain declining validity as grade level and writing complexity increase. Although transcription difficulties commonly limit writing quality in early writers (Berninger et al., 1997; McMaster, Ritchey, & Lembke, 2011), transcription is less of a limiting factor as students advance and composition length increases. In addition, upper elementary students' compositions exhibit more lexical diversity (Olinghouse & Graham, 2009), more syntactic complexity (e.g., longer sentences and more words per phrase; Beers & Nagy, 2011), better organization (Cox, Shanahan, & Sulzby, 1990; Galloway & Uccelli, 2015), and greater

differentiation by genre (Beers & Nagy, 2011). Broadening WE-CBM scoring to capture other aspects of quality, beyond transcription skills, may improve validity by strengthening construct representation as student grade level increases.

Automated Text Evaluation

Advances in the field of computational linguistics have resulted in various applications designed to generate quantitative indicators of text characteristics (e.g., Coh-Metrix; Graesser et al., 2014) that may improve assessment of student writing as it becomes more complex in upper elementary grades. In addition to descriptive metrics such as the total number of words, sentences, and paragraphs that are similar to some production-dependent WE-CBM metrics (e.g., TWW), programs like Coh-Metrix evaluate additional features of words, sentences, and discourse in compositions. For words, Coh-Metrix provides indicators of lexical diversity (e.g., the proportion of unique words in compositions), use of low-frequency words, relative frequencies of words classified by parts of speech, and psychological or semantic ratings of the words used, such as polysemy (the number of core meanings) and hypernymy (word specificity). For sentences in compositions, Coh-Metrix provides indicators of syntactic complexity, such as the average number of words before the main verb and number of words per noun phrase, and the density of specific syntactic patterns (e.g., incidence of noun and verb phrases and specific types of verb tenses). For characteristics of discourse, indicators capture semantic cohesion across sentences and paragraphs, referential cohesion (e.g., noun, pronoun, and content word overlap between sentences), and indicators of genre, such as narrativity, the extent to which the sample is similar to narrative texts; connectivity, the extent to which the sample contains connective words that describe relations among words and concepts; and temporality, which is

the extent to which the sample contains cues about temporal order of events and exhibits consistent usage of verb tenses.

These metrics were primarily developed as indicators of text reading comprehension difficulty (McNamara, Graesser, McCarthy, & Cai, 2014), but recent work demonstrates that they can be used to differentiate grade levels and predict quality judgments for essays written by high school and college students. For example, essays written by college students compared to ninth and eleventh grade students were rated higher on metrics assessing the number of words written, lexical diversity and word frequency, and syntactic complexity (Crossley, Weston, McLain Sullivan, & McNamara, 2011). Similarly, subsets of Coh-Metrix indicators were found to predict 42% of the variance in expert raters' holistic judgments (on a 6-point scale) of college student essay quality (McNamara, Crossley, & Roscoe, 2013). Less is known regarding the utility of Coh-Metrix to evaluate the writing skills of students in elementary grades, although one study (Puranik, Wagner, Kim, & Lopez, 2012) demonstrated differences in Coh-Metrix scores on writing samples from first and fourth grade students. This work suggests that Coh-Metrix scores capture grade-level differences in writing and predict judgments of writing quality in older students, but the extent to which Coh-Metrix scores can index general writing skill for universal screening and progress monitoring is unknown.

Purpose of the Current Study

There is a need to accurately and efficiently screen and monitor the writing skills of students with or at risk of LD; however, WE-CBM in its current form has limitations in technical adequacy, construct representation, and scoring feasibility as grade level increases. When evaluating the validity of alternative WE-CBM scoring methods, WE-CBM scores should both (a) correlate with rater judgements of writing quality on the samples used to generate the scores,

an important indicator of structural validity because scoring needs to adequately represent the construct of writing quality (Messick, 1995), and (b) correlate with quality judgements on other writing samples and standardized writing assessments to establish external validity (Messick, 1995) so that scores can index general writing skill (Deno, 1985). Thus far, research on automated text evaluation has primarily focused on structural validity, i.e., prediction of quality for the samples themselves, whereas WE-CBM research has primarily focused on external validity, e.g., prediction of performance on more comprehensive standardized writing assessments. For automated text evaluation to be useful for screening and progress monitoring within a CBM framework focused on indexing general writing skill, evidence of external validity is needed. Conversely, WE-CBM research would benefit from greater attention to structural validity given anecdotal evidence that teachers do not find commonly-used WE-CBM scoring metrics to adequately represent writing quality (Gansle et al., 2002; Ritchey & Coker, 2013).

The purpose of this study was to explore the validity of an automated text evaluation tool, specifically Coh-Metrix (Graesser et al., 2014), compared to traditional WE-CBM scoring, for use in evaluating elementary students' narrative writing samples. We compare the validity of Coh-Metrix relative to traditional WE-CBM, given that WE-CBM is commonly used in practice and has validity evidence for screening and progress monitoring, while addressing two research questions related to structural and external validity:

1) To what extent do Coh-Metrix scores, relative to WE-CBM scores, correlate with rater judgements of quality on scored writing samples?
2) To what extent do Coh-Metrix scores, relative to WE-CBM scores, serve as a general indicator of writing skill by correlating with rater judgements of quality on other writing samples?

Addressing these research questions will provide preliminary validity evidence for automated text evaluation for scoring writing samples within a WE-CBM framework; these results, in conjunction with future studies, have the potential to inform revised WE-CBM scoring procedures to better identify and monitor the progress of students with or at risk of LD in written expression. However, even with favorable preliminary validity evidence, more work will be necessary before automated text evaluation is ready to be implemented in schools for screening and progress monitoring. The current analyses address the specific Coh-Metrix scores and predictive models that may be useful in indexing general writing skill, but issues related to practical implementation such as how best to submit writing samples for analysis (e.g., students typing vs. handwriting samples, using handwriting recognition software) and how best to organize and present data to teachers for decision making will need to be addressed in future studies.

Method

Participants

Participants included 144 students in a suburban school district in the southwestern United States. Of the 144 students, 40 were in second grade, 37 in third grade, 37 in fourth grade, and 30 in fifth grade. Although the students varied in their exposure to the writing curriculum to some extent as a function of grade level, narrative writing was emphasized and taught at all grade levels. Participating students were 53% female, 49% White, 22% African American, 17% Hispanic, 8% Asian, and 4% identified as 2 or more races; 6% were English Language Learners, and 6% received special education services.

Procedures

After obtaining university ethics and school district research approval, parental consent for participation was elicited by sending letters home. Students with parental consent participated in the study. As part of data collection for a study on universal screening procedures (Keller-Margulis, Mercer, & Thomas, 2016), students completed three, seven-minute WE-CBM writing samples in November, February, and May of the same academic year. For the current study, student responses in November and February to one writing prompt ("I once had a magic pencil and…") that did not vary across grades were analyzed. Procedural fidelity data were collected for approximately 89% of all writing sample administrations in the original study with an average of 99% administration steps successfully completed (Keller-Margulis et al., 2016). The study authors and other graduate students in school psychology evaluated the writing samples.

Measures

As detailed subsequently, writing samples were evaluated for overall writing quality (new in this study) and scored for traditional WE-CBM indicators (reported in Keller-Margulis et al., 2016) and with Coh-Metrix (new in this study).

Writing quality. Prior studies on WE-CBM and automated text evaluation have often used holistic ratings (e.g., a 7-point quality scale) or analytic rubrics to assess writing quality for convergent validity; however, both methods have been criticized due to concerns with inter-rater reliability and limited variability in scores across writing samples (Allen, Poch, & Lembke, 2018; Gansle, VanDerHeyden, Noell, Resetar, & Williams, 2006). To address these concerns with inter-rater reliability and limited variability, we evaluated holistic writing quality, considering idea development and organization of ideas, using the method of paired comparisons (Thurstone, 1927). In this method, raters are repeatedly presented with pairs of samples and then indicate which sample in the pair is of better overall quality. Compared to assigning a particular

quality score, e.g., evaluating whether a writing sample should be scored as a 4 vs. 5 on a 7-point quality scale, it is cognitively easier for raters to determine whether a particular sample is of better or worse quality than another writing sample (Heldsinger & Humphry, 2010). Once a large number of writing sample pairs were evaluated by the raters, these judgements were submitted to an algorithm[1] (Jiang, Lim, Yao, & Ye, 2011) that assigns each writing sample a quality score ranging from -1 to +1 that represents the tendency of the sample to be rated as better than other samples in the set.

The optimal number of paired comparisons for reliable estimates of overall quality was determined by investigating the stability of the algorithm-generated quality scores as groups of 500 comparisons were added to the total comparisons until there was minimal change in the quality scores. For the fall samples, 8,000 pairs of samples were initially evaluated—quality scores based on 8,000 and 7,500 evaluated pairs were highly stable, with a concordance correlation coefficient (Lin, 1989), a measure of absolute agreement, of 1.00 between the scores based on 8,000 and 7,500 evaluated pairs. Because these initial analyses also indicated that quality scores were stable with much fewer than 8,000 evaluated pairs, fewer pairs of winter samples were evaluated, and stability of quality scores was evident with 3,000 evaluated paired samples. The concordance correlation coefficient was 1.00 between quality scores based on 3,000 and 2,500 evaluated pairs. In sum, these analyses provide evidence that the algorithm-generated quality scores are reliable, specifically that enough paired comparisons were

[1] To describe the algorithm, let Y_{ij} be equal to +1 when a rater prefers sample *i* over sample *j*, or -1 when *j* is preferred over *i*. The algorithm then identifies a set of quality scores (x) for the samples that represent the tendency for sample *i* be preferred over sample *j*, i.e., if $x_i > x_j$, then sample *i* is likely to be rated better than sample *j*. The quality scores are optimized to minimize the squared loss function, $L(x_i - x_j, Y_{ij})$; the quality scores range from approximately -1 to +1 due to the coding of Y_{ij} as -1 or +1.

conducted such that additional comparisons were unlikely to substantively change the quality scores.

Traditional WE-CBM. The writing samples were scored for six metrics that are commonly used in practice and/or have the most evidence of validity in WE-CBM research (Hosp, Hosp, & Howell, 2016; McMaster & Espin, 2007; Romig et al., 2016). We scored three metrics, Total Words Written (TWW), Words Spelled Correctly (WSC), and Correct Word Sequences (CWS), that were used to derive three additional metrics: Percentage of Words Spelled Correctly (%WSC), Correct Minus Incorrect Word Sequences (CIWS), and Percentage Correct Word Sequences (%CWS). TWW is a count of the total words in the composition, including misspelled and nonsense words. WSC is a count of the number of words spelled correctly when considered in isolation. CWS is a count of neighboring units (i.e., word-word, punctuation-word, word-punctuation) with correct spelling, punctuation, and grammar that make sense in the context of the sentence. All raters attended training and completed practice scoring with the requirement that they reach 90% reliability before participating. Inter-rater reliability, based on 20% of samples that were scored by two raters, was high, with concordance correlation coefficients (Lin, 1989) between .99 and 1.00 for TWW, WSC, and CWS, and .92 for IWS. More detailed reliability information for the WE-CBM scores is presented in Keller-Margulis et al. (2016).

Coh-Metrix. The samples were computer-scored using Coh-Metrix (McNamara et al., 2014). To enter the writing samples into Coh-Metrix, hand written samples were typed by a graduate student in school psychology. The typed samples were then independently checked by another graduate student for accuracy and discrepancies were resolved before entry.

Because we had limited *a priori* information to determine which metrics would perform best as indicators of writing skill, all of the metrics generated by Coh-Metrix were considered for inclusion in the predictive models (98 after removal of redundant metrics), with the best-performing metrics empirically selected through the model building process, as detailed in the Data Analysis section. The metrics capture aspects of the words used in the samples, such as type-token ratio, a measure of lexical diversity operationalized as the proportion of words in a sample that are unique; average word frequency of all words, which captures the extent to which high vs. low frequency vocabulary words are used; and average word polysemy, which is a measure of how many meanings the words have as a measure of vocabulary specificity. The metrics also quantify aspects of the sentences generated in the samples, such as the mean number of words before the main verb and the average number of modifiers per noun phrase as measures of syntactic complexity. Last, the metrics capture aspects of discourse, such as referential cohesion, including the extent to which adjacent sentences overlap in nouns, arguments, and content words; narrativity, the extent to which the sample is similar to narrative texts; connectivity, the extent to which the sample contains connective words that describe relations among words and concepts; and temporality, which is the extent to which the sample contains cues about temporal event order and exhibits consistent usage of verb tenses.

Data Analysis

As preliminary analyses, we first evaluated the extent to which there were between-grade differences in fall and winter writing quality in one-way analyses of variance (ANOVA) with grade level as a factor. We conducted separate ANOVAs for fall and winter, rather than a mixed ANOVA with time as an additional factor, because samples were rated relative to each other

within each time point; thus, no overall differences by time were expected. As an indicator of effect size and the proportion of total variance that is between- vs. within-grades, η^2 is reported.

Our main analyses were conducted within an applied predictive modeling framework (Kuhn & Johnson, 2013), in which the primary goal is to build a model using training data that can accurately predict scores on untrained, test data. Unlike traditional applications of linear regression in which minimizing bias between model predictions and training data is the main concern, in applied predictive modeling, the error for model predictions on untrained, test data is a key concern. Specifically, we identified models that fit training data well enough, but not so well that overfitting to the training data would occur, so that the model could be used to generate accurate predictions on untrained, test data sets. In our analyses, we first trained models with Coh-Metrix scores and WE-CBM scores as predictors of holistic quality ratings on the samples themselves (e.g., Coh-Metrix scores on fall samples predicting fall quality ratings), and then tested the performance of the trained models on other writing samples (e.g., fall models applied to winter sample data to predict winter quality ratings).

In applied predictive modeling, many different prediction algorithms are available. Because we had a large number of predictors relative to the number of writing samples to be analyzed, we focused on algorithms that either explicitly select predictors (i.e., include or exclude predictors) or implicitly select predictors (e.g., down weight less informative predictors). All of the selected algorithms can handle high-dimensional problems where number of predictors exceeds sample size (Hastie, Tibshirani, & Friedman, 2009). Although the number of predictors relative to the number of writing samples would be non-optimal in traditional multiple regression where the focus is on the statistical significance of individual predictors, our purpose was to apply all useful information on the predictors, as empirically determined in the model training

process, to generate model-predicted quality scores to be evaluated with test data in subsequent analyses (e.g., correlations with writing quality on other samples).

The following algorithms were evaluated: (a) best-subset multiple regression using forward selection of predictors, in which predictors are added sequentially based on potential improvement in model fit; (b) Bayesian lasso regression, in which predictors are weighted by shrinking regression coefficients and some predictors are removed by requiring the sum of the absolute values of the regression coefficients to be less than a specific value; (c) elastic-net regression, which weights and selects predictors similarly to lasso regression but adds a second shrinkage penalty based on squared regression coefficients, not just the sum of absolute regression coefficients; (d) bagged multivariate adaptive regression splines (MARS), a non-parametric regression approach that can handle non-linearities and interactions among predictors in which regression terms, consisting of piecewise linear functions, and the products of regression terms already in the model, are added in a forward selection process, with final predictions based on averaged results over multiple models; (e) gradient boosted regression trees, another non-parametric approach in which regression trees (i.e., successive splits of data into regions at values of specific predictors that minimize prediction error) are added to the model in a forward stagewise process to further minimize residuals from prior trees in the model; (f) random forest regression, another non-parametric approach in which regression trees are built by randomly selecting subsets of predictors for consideration for each split in the tree and then averaging the trees in an ensemble model; and (g) partial least squares regression, in which linear combinations of the predictors are identified that maximize both variance explained in the predictors and in the criterion variable, in contrast to multiple linear regression in which only

variance explained in the criterion is maximized. Detailed descriptions of these algorithms are presented in Hastie et al. (2009).

In the model training process, (a) WE-CBM and then (b) Coh-Metrix scores were entered as potential predictors of quality ratings, e.g., fall WE-CBM and then fall Coh-Metrix scores as predictors of fall quality ratings. Most of the algorithms have one or more tuning parameters that need to be optimized, such as the number of predictor variables considered at each step of building trees in random forest regression. To determine optimal values of these tuning parameters, models were fit with adaptive resampling of better-performing tuning parameters using repeated four-fold cross-validation (Hastie et al., 2009). Specifically, the training data were randomly divided into four equal folds, with three of the folds used to build models and a fourth used to calculate root mean square error (RMSE) of prediction (a validation fold) until all folds have served as a validation fold, with the process repeated 10 times to yield aggregated RMSE values across different tuning parameter values for each algorithm. After optimal tuning parameters were identified, a final round of repeated (2,500 times) four-fold cross validation was performed to enable between-algorithm comparisons of RMSE values based on the same 10,000 resampled training data sets. This process is fully automated in the *caret* package (Kuhn, 2016) in R (R Core Team, 2017).

These RMSE values were used to identify well-performing models to evaluate on untrained, test data sets. Best-subset multiple regression, regardless of RMSE, was selected for interpretability because the relative importance of each predictor is readily ascertained based on standardized beta coefficients; in addition, one other algorithm was selected based on smallest RMSE for both the WE-CBM and Coh-Metrix predictor models in fall and winter. The models' performance on test data were evaluated in two ways: by determining the extent to which the (a)

model-predicted quality ratings, based on the same training data used to build the model, correlated with writing sample quality ratings for the same students at a different time point (e.g., correlating predicted fall quality scores based on fall model and fall writing samples with winter quality ratings) and (b) model-predicted quality ratings, when based on writing sample data not used to build the model, correlated with actual quality ratings (e.g., correlating predicted winter quality scores based on fall model applied to winter sample data with winter quality ratings). For these final analyses, there were some missing data because 16% of the sample completed writing samples at only the fall or winter assessment periods; thus, multiple imputation (with 5,000 datasets) was used to appropriately handle missing data (Baraldi & Enders, 2010) when calculating the correlations. To aid in the interpretation of validity coefficients, we used descriptive labels similar to those of the McMaster and Espin (2007) review of WE-CBM research: $r \geq .80$, relatively strong; $r = .70$ to .79, moderately strong; $r = .60$ to .69, moderate; and $r < .60$, weak.

Results

Means and standard deviations for writing quality scores by grade level and time are presented in Table 1. There were statistically significant differences in writing quality in fall, F (3, 129) = 43.05, $p < .001$, $\eta^2 = .50$, and winter, F (3, 117) = 30.37, $p < .001$, $\eta^2 = .44$, with a general trend of increasing average quality across grade levels. Results of pairwise tests of mean differences by grade also are presented in Table 1. Approximately 50% and 56% of the total variance in fall and winter writing quality was within grade levels (1- η^2).

Model Training

Variance explained (R^2) in writing quality ratings by prediction algorithm and time point is presented in Table 2. For these models, WE-CBM and then Coh-Metrix scores on samples at

each time point were entered as predictors of quality ratings on the same samples at the same time point (i.e., training data sets). For WE-CBM, there were small differences in resampled R^2 values across the algorithms at fall (R^2 = .686 to .702); predicted quality values (fall quality based on fall data and model, and winter quality based on winter data and fall model) were generated for bagged MARS as the best-performing algorithm and best subset regression as an easily interpretable algorithm for relative predictor importance. For WE-CBM at winter, R^2 values were lower (.539 to .585) compared to fall; predicted quality values were generated for best subset regression as the best-performing algorithm, and elastic-net regression as a second-best comparison algorithm (R^2 = .583). For Coh-Metrix at fall, best subset regression was selected as the best-performing algorithm (R^2 = .771), and Bayesian lasso regression was selected as a second-best comparison algorithm (R^2 = .730). For Coh-Metrix at winter, gradient boosted regression trees were selected as the best algorithm (R^2 = .651), with best subset regression also selected for interpretation (R^2 = .618). At both fall and winter on training data, the best performing Coh-Metrix algorithms outperformed the best performing WE-CBM algorithms, R^2 = .771 vs .701 at fall and R^2 = .651 vs. .585 at winter; tests of differences in dependent correlations between predicted quality and evaluated quality for Coh-Metrix and WE-CBM were all $p < .05$ in favor of Coh-Metrix. Overall, these results provide evidence of structural validity for the Coh-Metrix and WE-CBM scores, i.e., that they capture substantive aspects of writing quality on the samples themselves; however, these difference between Coh-Metrix and WE-CBM models in R^2 are more meaningful if replicated on test data, which would provide evidence of incremental external validity.

Relative Predictor Importance

To aid in the interpretation of which Coh-Metrix and WE-CBM scores contributed to quality predictions on the training data, standardized beta coefficients for the predictors included in the best subset regression models are presented in Table 3. In the Coh-Metrix models, DESWC (Descriptives: Word Count) accounted for roughly half of the variance explained in writing quality (β = .690 and .701, for fall and winter, respectively). By contrast, WSC was the strongest predictor (β = .745) in the fall WE-CBM model, and CWS was the strongest predictor (β = .975) in the winter model, with a caveat that the strong correlation between CWS and CIWS (*r* = .936) contributed to multicollinearity that complicates interpretation of individual predictors in that model. Although WSC and CWS, instead of TWW, were included in the WE-CBM models based on forward selection, it is important to note that WSC and CWS were highly correlated with TWW (*r* = .985 and .914, respectively), thus, these metrics largely reflected word count. In addition to word count, the average number of letters in words was also included in the fall and winter Coh-Metrix models. In sum, for both WE-CBM and Coh-Metrix, the total number of words in student compositions was most crucial in predicting judgements of writing quality.

Model Evaluation with Test Data

The following analyses address external validity, i.e., the extent to which model-predicted quality scores correlate with writing quality on other samples. Correlations of model-predicted quality scores and rated writing quality are presented in Table 4. The correlations are presented in three groups that differed in the procedures used to generate the predicted values: (a) fall sample data and the fall model for predicted fall quality, (b) winter sample data and winter model for predicted winter quality, and (c) fall model and winter sample data for predicted winter quality. Correlations that are bolded are tests of external validity through cross-validation, i.e., when model-predicted quality was correlated with rated quality on samples not

used to train the model or when quality predictions were generated from a model based on writing sample data not used to train the model.

In general, three main patterns are evident in the correlations: (a) correlations were smaller with test compared to training data (i.e., smaller external than structural validity coefficients), (b) all test data correlations were moderately to relatively strong (r = .730 to .807), and (c) differences among test data correlations with WE-CBM vs. Coh-Metrix as predictors and across predictive algorithms were minimal. Although test data correlations with best-subset regression vs. alternative algorithms were quite similar, test data correlations for the alternative algorithms were larger, albeit modestly so (Δr = .001 to .048), for pairs of models with the same input scores.

Discussion

Technically adequate writing screening measures that can efficiently identify and monitor progress for upper elementary students with or at-risk for LD in written expression are greatly needed. Construct underrepresentation in traditional WE-CBM scoring may contribute to declining validity coefficients as student grade level increases (McMaster & Espin, 2007), and the trend in WE-CBM research toward more complex scoring procedures to improve validity may reduce scoring feasibility (Espin et al., 1999; Gansle et al., 2002), particularly with longer duration samples and more than one sample administered per occasion. Findings from the current study address key issues related to structural and external validity.

First, when predicting rated quality on the training data as evidence of structural validity, composites of traditional WE-CBM and Coh-Metrix scores, collectively, were moderately to relatively strong (r = .768 to .887 for the best-subset algorithm) and higher than nearly all of the validity coefficients at comparable grade levels for WE-CBM scores with holistic or analytic

quality ratings summarized in the McMaster and Espin (2007) review. For example, WE-CBM scores on 6 min story samples for second through fifth grade students were weakly to moderately correlated with 7-point holistic ratings of the same samples at r = .36 to .70 depending on the specific WE-CBM score and grade level (Parker, Tindal, & Hasbrouck, 1991). Similarly, WE-CBM scores on 3 min story samples for second and fourth grade were weakly correlated with analytic ratings of quality on the same samples at r = .34 to .58 (Jewell & Malecki, 2005). The higher validity coefficients in the current study are likely due to several factors. In the current study, validity coefficients were calculated across grades; by contrast, within-grade validity coefficients were reported in most of the reviewed studies addressing structural validity in McMaster and Espin (2007). Although within-grade variability contributed roughly 50% of the variance in writing quality scores in the current study, the greater total variability by including between-grade variance likely contributed to larger validity coefficients. Also, we evaluated composite scores instead of individual WE-CBM metrics, and prior research indicates that using composite WE-CBM scores, although uncommon in practice, can improve convergent validity (Codding et al., 2015; Espin et al., 1999). Last, prior WE-CBM studies with 5- or 7- point holistic quality ratings as a validation measure may have attenuated validity coefficients due to the ordinal response format, restriction of range, and non-optimal interscorer reliability (Gansle et al., 2006; Zumbo, Gadermann, & Zeisser, 2007). By contrast, the paired comparison method used in the current study for evaluating quality yields greater data variability compared to ratings with a fixed number of options, and may also improve interscorer reliability to some extent.

Second, the improved structural validity coefficients, in comparison to many of those reported in McMaster and Espin (2007) for WE-CBM scores and holistic or analytic quality ratings of the same samples, were also evident in cross-validation (r = .730 to .807), providing

some external validity evidence in applications where the WE-CBM and Coh-Metrix scores, predictive models, and criterion quality ratings were not all based on the same writing samples. The magnitude of these validity coefficients is notable given that they were quite comparable to the correlation between evaluated quality on the fall and winter samples (r = .800) and logically should not exceed this value. Although uncommon in WE-CBM research, true cross-validation analyses, beyond the resampling-based cross-validation used during initial model fitting, are particularly important in applied predictive modeling given that nearly perfect correlations can be obtained between model predictions and evaluated quality on training samples by overfitting models to the training data (Hastie et al., 2009). The cross-validation analyses demonstrate that we did not overfit models to the training data and provide evidence of the potential of applied predictive modeling to generate predicted quality scores that can serve as general indicators of writing skill.

Third, we found minimal differences in external validity coefficients across predictive algorithms; however, performance across different algorithms should continue to be examined. When building the predictive models, metrics reflecting word count were heavily weighted, thus one predictor disproportionately contributed to model-predicted quality scores. This finding is not unique to the current study; indeed, overreliance on composition length is a common criticism of commercial automated text evaluation programs that are currently used in high-stakes assessments (Perelman, 2014). In addition, WE-CBM metrics representing or highly correlated with word count have long been studied as indicators of general writing skill (McMaster & Espin, 2007; Romig et al., 2016). It is possible that the short task duration (7 min) constrained student ability to plan, organize, and revise, thereby reducing overall writing quality and the need for more complex scoring metrics to predict it. With longer task durations, as has

been recommended in WE-CBM research to improve technical adequacy (e.g., Espin et al., 2008; Keller-Margulis et al., 2016), or with other types of writing prompts than story, e.g., informational, it is possible that word count would be a less robust predictor or that there would be non-linearities or complex patterns of interactions among predictors that would boost performance of alternative algorithms compared to best-subset regression (Hastie et al., 2009).

Fourth, although we predicted that the greater range of text characteristics scored in Coh-Metrix compared to traditional WE-CBM would improve representation of the writing quality construct and, in turn, yield higher validity coefficients for Coh-Metrix score models, this prediction did not fully hold. We obtained higher structural validity coefficients on the training samples for Coh-Metrix compared to traditional WE-CBM models; however, no substantive differences in external validity coefficients were evident between the models during cross-validation with test data. Notably, the higher structural validity coefficients for Coh-Metrix indicate that scores better represented evaluated writing quality on the samples themselves; although WE-CBM research has largely focused on external validity through prediction of criterion measures, continued investigation of structural validity is important to address anecdotal reports that teachers perceive WE-CBM scoring to insufficiently represent writing quality (Gansle et al., 2002; Ritchey & Coker, 2013). Given the issues raised above with the short task duration and use of only one writing prompt (and prompt genre) in this study, additional research is needed before fully dismissing the potential of automated text evaluation to improve construct representation and validity.

Limitations and Future Directions

These findings should be considered in light of several limitations. First, the sample size was too small to permit separate analyses by grade, thus, future studies with larger samples that

would permit such analyses are recommended to determine the extent to which specific predictors of writing quality vary by grade level. Second, future studies would benefit from using longer duration samples to improve reliability and validity of WE-CBM and to potentially improve the performance of more complex Coh-Metrix indicators that are sensitive to composition length. Third, we only examined narrative writing samples, but future research should examine informational and argumentative genres that are emphasized in curricula (National Governors Association Center for Best Practices, 2010). Last, although we checked cross-validation with samples administered on different occasions, an extension of prior WE-CBM studies using holistic and analytic ratings of only the scored samples themselves, future research would benefit from inclusion of more comprehensive standardized writing assessments as external validity measures of general writing skill.

Practical Implications

The results of this study potentially have implications for the screening and progress monitoring of students with or at risk of LD in written expression in upper elementary grades. WE-CBM can be used to efficiently collect data on overall student writing performance that can be used for decision making about student risk status and progress during instruction and intervention. Such assessments are not intended to provide detailed diagnostic information about specific areas in need of improvement, and we believe that approaches like WE-CBM for decisions about overall performance are best combined with detailed qualitative feedback from teachers about specific aspects of composition in need of improvement to realize the gains in achievement associated with formative assessment (Graham, Hebert, & Harris, 2015). Although we focused on the use of automated text evaluation within a WE-CBM framework to index overall writing skill, other research demonstrates the benefits of automated text evaluation to

provide more detailed diagnostic feedback to students (Roscoe & McNamara, 2013). Evidence suggests that teachers provide more feedback about higher-level writing skills when feedback from automated text evaluation is also provided to students (Wilson & Czik, 2016).

Given the identified limitations of traditional WE-CBM in technical adequacy and scoring feasibility, other ways to efficiently and defensibly score and interpret student writing samples are greatly needed. The current study suggests that Coh-Metrix can be used for computer scoring WE-CBM writing samples with potential gains in feasibility, plus potentially fewer concerns with monitoring and maintaining inter-scorer agreement across multiple raters. The comparable external validity coefficients for Coh-Metrix vs. WE-CBM suggest that automated text evaluation can potentially replace hand scored WE-CBM metrics without compromising data quality; however, these results need to be confirmed with more comprehensive external validity measures and the classification accuracy of decisions based on automated scoring needs to be evaluated before recommending its use for screening or progress monitoring (Smolkowski, Cummings, & Strycker, 2016).

Regarding feasibility, although we did not record the time required to hand score the WE-CBM metrics and verify inter-scorer agreement in the current study, prior studies have estimated that it requires four to five minutes to score for multiple metrics per student, depending on the number of specific metrics scored (Espin et al., 1999). These time estimates are likely to be higher with the longer duration writing samples and multiple writing samples needed for reliable estimates of student writing skill (Graham, Hebert, Sandbank, & Harris, 2016; Keller-Margulis et al., 2016). These feasibility issues compound when conducting universal screening of all students in a class or school, underscoring the need for more feasible ways to score WE-CBM.

Similarly, although we transcribed handwritten student writing samples for entry in Coh-Metrix in the current study, this potential feasibility limitation may be lessened as keyboarding is increasingly used by students for composition and with the rapid development of neural network models for computerized handwriting recognition (Doetsch, Kozielski, & Ney, 2014). Ultimately, if future research continues to identify benefits for automated text evaluation for universal screening and progress monitoring within a CBM framework, several specific issues will need to be resolved: (a) writing samples will need to be easily submitted for analysis by having students type compositions or through automated handwriting recognition, (b) models such as the ones trained in the current study will need to be implemented automatically by software to generate predicted quality scores, and (c) software will need to facilitate data-based decision making by simplifying data display and analysis for individuals and groups of students. Before addressing these practical concerns, however, additional research on automated text evaluation is needed with longer duration and multiple-genre samples and with more robust writing assessments to establish external validity. Although preliminary, we hope that the findings from the current study contribute to ongoing efforts to revise WE-CBM administration and scoring procedures to efficiently yield defensible data for use in screening and progress monitoring of students with or at risk of LD in written expression in upper elementary grades and beyond.

References

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

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology, 48,* 5-37. doi:10.1016/j.jsp.2009.10.001

Beers, S. F., & Nagy, W. E. (2011). Writing development in four genres from grades three to seven: Syntactic complexity and genre differentiation. *Reading and Writing, 24,* 183-202. doi:10.1007/s11145-010-9264-9

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

Berninger, V. W., Vaughan, K. B., Abbott, R. D., Abbott, S. P., Rogan, L. W., Brooks, A., . . . Graham, S. (1997). Treatment of handwriting problems in beginning writers: Transfer from handwriting to composition. *Journal of Educational Psychology, 89,* 652-666. doi:10.1037/0022-0663.89.4.652

Codding, R. S., Petscher, Y., & Truckenmiller, A. (2015). CBM reading, mathematics, and written expression at the secondary level: Examining latent composite relations among indices and unique predictions with a state achievement test. *Journal of Educational Psychology, 107,* 437-450. doi:10.1037/a0037520

Cox, B. E., Shanahan, T., & Sulzby, E. (1990). Good and poor elementary readers' use of cohesion in writing. *Reading Research Quarterly, 25,* 47-65. doi:10.2307/747987

Crossley, S. A., Weston, J. L., McLain Sullivan, S. T., & McNamara, D. S. (2011). The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication, 28*, 282-311. doi:10.1177/0741088311410188

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

Deno, S. L., Marston, D., & Mirkin, P. (1982). Valid measurement procedures for continuous evaluation of written expression. *Exceptional Children, 48*, 368-371. doi:10.1177/001440298204800417

Doetsch, P., Kozielski, M., & Ney, H. (2014). *Fast and robust training of recurrent neural networks for offline handwriting recognition.* Paper presented at the 14th International Conference on Frontiers in Handwriting Recognition, Heraklion, Greece. doi:10.1109/ICFHR.2014.54

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, 15*, 5-27. doi:10.1080/105735699278279

Espin, C. A., 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. doi:10.1177/002246690003400303

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

Galloway, E. P., & Uccelli, P. (2015). Modeling the relationship between lexico-grammatical and discourse organization skills in middle grade writers: Insights into later productive language skills that support academic writing. *Reading and Writing, 28*, 797-828. doi:10.1007/s11145-015-9550-7

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.

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

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. doi:10.1086/678293

Graham, S., Hebert, M., & Harris, K. R. (2015). Formative assessment and writing: A meta-analysis. *The Elementary School Journal, 115*, 523-547. doi:10.1086/681947

Graham, S., Hebert, M., Sandbank, M. P., & Harris, K. R. (2016). Assessing the writing achievement of young struggling writers: Application of Generalizability Theory. *Learning Disability Quarterly, 39*, 72-82. doi:10.1177/0731948714555019

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

Heldsinger, S., & Humphry, S. (2010). Using the method of pairwise comparison to obtain reliable teacher assessments. *The Australian Educational Researcher, 37*, 1-19. doi:10.1007/BF03216919

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

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

Jiang, X., Lim, L.-H., Yao, Y., & Ye, Y. (2011). Statistical ranking and combinatorial Hodge theory. *Mathematical Programming, 127*, 203-244. doi:10.1007/s10107-010-0419-x

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. doi:10.1037/spq0000126

Kuhn, M. (2016). *caret: Classification and Regression training. R package version 6.0-70.* Retrieved from https://CRAN.R-project.org/package=caret

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. New York: Springer.

Lin, L. I. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics, 45*, 255-268. doi:10.2307/2532051

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. doi:10.1002/pits.10096

Marston, D., & Deno, S. (1981). *The reliability of simple, direct measures of written expression*. University of Minnesota, Institute for Research on Learning Disabilities. Minneapolis, MN.

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

McMaster, K. L., Ritchey, K. D., & Lembke, E. (2011). Curriculum-based measurement for beginning writers: Recent developments and future directions. In T. E. Scruggs & M. A. Mastropieri (Eds.), *Assessment and intervention: Advances in learning and behavioral disabilities* (Vol. 24, pp. 111-148). Bingley, UK: Emerald Group Publishing Limited.

McNamara, D. S., Crossley, S. A., & Roscoe, R. (2013). Natural language processing in an intelligent writing strategy tutoring system. *Behavior Research Methods, 45*, 499-515. doi:10.3758/s13428-012-0258-1

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. New York, NY: Cambridge University Press.

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*, 85-95. doi:10.1037/a0029123

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

National Center for Education Statistics. (2012). *The nation's report card: Writing 2011*. Retrieved from http://www.nationsreportcard.gov/writing_2011/

National Governors Association Center for Best Practices. (2010). *Common core State Standards for English Language Arts.* Retrieved from http://www.corestandards.org

Olinghouse, N. G., & Graham, S. (2009). The relationship between the discourse knowledge and the writing performance of elementary-grade students. *Journal of Educational Psychology, 101,* 37-50. doi:10.1037/a0013462

Parker, R., Tindal, G., & Hasbrouck, J. (1991). Countable indices of writing quality: Their suitability for screening-eligibility decisions. *Exceptionality, 2,* 1-17. doi:10.1080/09362839109524763

Perelman, L. (2014). When "the state of the art" is counting words. *Assessing Writing, 21,* 104-111. doi:10.1016/j.asw.2014.05.001

Puranik, C. S., Wagner, R. K., Kim, Y.-S., & Lopez, D. (2012). Multivariate assessment of processes in elementary students' written translation. In M. Fayol, D. Alamargot, & V. W. Berninger (Eds.), *Translation of thought to written text while composing: Advancing theory, knowledge, research methods, tools, and applications.* (pp. 249-274). NY: Psychology Press.

R Core Team. (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

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,* 89-119. doi:10.1080/10573569.2013.741957

Romig, J. E., Therrien, W. J., & Lloyd, J. W. (2016). Meta-analysis of criterion validity for curriculum-based measurement in written language. *Journal of Special Education, 51,* 72-82. doi:10.1177/0022466916670637

Roscoe, R. D., & McNamara, D. S. (2013). Writing pal: Feasibility of an intelligent writing strategy tutor in the high school classroom. *Journal of Educational Psychology, 105,* 1010-1025. doi:10.1037/a0032340

Salahu-Din, D., Persky, H., & Miller, J. (2007). *The nation's report card: Writing*. Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Silberglitt, B., Parker, D., & Muyskens, P. (2016). Assessment: Periodic assessment to monitor progress. In S. R. Jimerson, M. K. Burns, & A. M. VanDerHeyden (Eds.), *Handbook of response to intervention: The science and practice of multi-tiered systems of support* (2nd ed., pp. 271-292). New York: Springer.

Smolkowski, K., Cummings, K. D., & Strycker, L. (2016). An introduction to the statistical evaluation of fluency measures with signal detection theory. In K. D. Cummings & Y. Petscher (Eds.), *The fluency construct: Curriculum-based measurement concepts and applications* (pp. 187-222). New York: Springer.

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

Weissenburger, J. W., & Espin, C. A. (2005). Curriculum-based measures of writing across grade levels. *Journal of School Psychology, 43,* 153-169. doi:10.1016/j.jsp.2005.03.002

Wilson, J., & Czik, A. (2016). Automated essay evaluation software in English Language Arts classrooms: Effects on teacher feedback, student motivation, and writing quality. *Computers & Education, 100,* 94-109. doi:10.1016/j.compedu.2016.05.004

Zumbo, B. D., Gadermann, A. M., & Zeisser, C. (2007). Ordinal versions of coefficients alpha and theta for Likert rating scales. *Journal of Modern Applied Statistical Methods, 6*, 21-29. doi:10.22237/jmasm/1177992180

Table 1

Mean Writing Quality Ratings in Fall and Winter by Grade

	Fall			Winter		
Grade	*M*	*SD*	*n*	*M*	*SD*	*n*
2nd	-.58[a]	.31	37	-.51[a]	.32	31
3rd	.08[b]	.42	31	.11[b]	.44	32
4th	.16[b]	.46	35	.25[b]	.40	30
5th	.47[c]	.39	30	.36[b]	.38	28
Overall	.00	.55	133	.04	.51	121

Note. Grade-level means with different letter superscripts ([abc]) are statistically different based on Scheffé tests at $p < .05$.

Table 2

Variance Explained (R^2) by Predictive Algorithm and Time based on Repeated 4-fold Cross-validation with Training Data

	Fall		Winter	
Algorithm	WE-CBM	Coh-Metrix	WE-CBM	Coh-Metrix
Best Subset Regression	.689	**.771**	**.585**	.618
Bayesian Lasso	.686	.730	.578	.565
Elastic-Net Regression	.687	.724	.582	.589
Bagged MARS	**.702**	.703	.581	.611
Gradient Boosted Trees	.696	.690	.539	**.651**
Random Forest	.694	.682	.562	.632
Partial Least Squares	.686	.686	.577	.539

Note. Fall n = 133, Winter n = 131. WE-CBM = written expression curriculum-based measurement, MARS = multivariate adaptive regression splines. The largest R^2 values by predictor type (WE-CBM or Coh-Metrix) and time point are bolded.

Table 3

Best-fitting Models in Fall and Winter for Best Subset Regression using Forward Selection

		Fall			Winter		
Model	Predictor	β	*p*	R^2	β	*p*	R^2
WE-CBM	WSC	.745	<.001	.689			.585
	%CWS	.131	.031				
	CWS				.975	<.001	
	CIWS				-.228	.179	
Coh-Metrix	DESWC	.690	<.001	.771	.702	<.001	.618
	DESWLlt	.208	<.001		.319	<.001	
	WRDHYPn	.239	<.001				
	LDMTLD	.108	.062				
	WRDPRP2	-.127	.003				
	WRDFRQc	.116	.014				
	LDTTRc	.130	.015				
	SMINTEp	-.095	.034				

Note. Fall *n* = 133, Winter *n* = 131. WE-CBM = written expression curriculum-based measurement, WSC = words spelled correctly, %CWS = percentage correct word sequences, CWS = correct word sequences, CIWS = correct minus incorrect word sequences, DESWC = Descriptive: word count, DESWLlt = Descriptive: word length (average number of letters), WRDHYPn = Word information: mean hypernymy values for nouns, LDMTLD = Lexical diversity: Measure of Textual Lexical Diversity, WRDPRP2 = Word information: second-person pronoun incidence, WRDFRQc = Word information: mean CELEX word frequency for content words, LDTTRc = Lexical diversity: type-token ratio for content words, SMINTEp = Situation model: intentional verbs incidence.

Table 4

Correlations of Model-Predicted Writing Quality with Rated Writing Quality

Procedure	Input Scores	Algorithm	Fall Quality *r*	Winter Quality *r*
Predicted fall quality based on fall model	Fall WE-CBM	Best Subset	.828	**.741**
		Bagged MARS	.854	**.767**
	Fall Coh-Metrix	Best Subset	.887	**.745**
		Bayesian Lasso	.912	**.766**
Predicted winter quality based on winter model	Winter WE-CBM	Best Subset	**.755**	.768
		Elastic-Net	**.756**	.767
	Winter Coh-Metrix	Best Subset	**.759**	.812
		Boosted Trees	**.807**	.991
Predicted winter quality based on fall model	Winter WE-CBM	Best Subset	**.750**	**.755**
		Bagged MARS	**.762**	**.772**
	Winter Coh-Metrix	Best Subset	**.730**	**.734**
		Bayesian Lasso	**.754**	**.758**

Note. n = 144. WE-CBM = written expression curriculum-based measurement, MARS = multivariate adaptive regression splines. Bolded values are test-data correlations involving model-predicted and rated writing quality scores when the predicted quality scores and rated quality scores were from different samples and/or the data used to train the model and generate predicted values differed.