



# Cost analysis and cost-effectiveness of hand-scored and automated approaches to writing screening

Michael Matta<sup>a,\*</sup>, Milena A. Keller-Margulis<sup>a</sup>, Sterett H. Mercer<sup>b</sup>

<sup>a</sup> Department of Psychological, Health & Learning Sciences, University of Houston, USA

<sup>b</sup> Department of Educational and Counselling Psychology, and Special Education, The University of British Columbia, Canada

## ARTICLE INFO

Editor: Craig A. Albers

### Keywords:

Written expression  
Curriculum-based measurement  
Automated text evaluation  
Universal screening  
Cost analysis  
Cost effectiveness

## ABSTRACT

Although researchers have investigated technical adequacy and usability of written-expression curriculum-based measures (WE-CBM), the economic implications of different scoring approaches have largely been ignored. The absence of such knowledge can undermine the effective allocation of resources and lead to the adoption of suboptimal measures for the identification of students at risk for poor writing outcomes. Therefore, we used the Ingredients Method to compare implementation costs and cost-effectiveness of hand-calculated and automated scoring approaches. Data analyses were conducted on secondary data from a study that evaluated predictive validity and diagnostic accuracy of quantitative approaches for scoring WE-CBM samples. Findings showed that automated approaches offered more economic solutions than hand-calculated methods; for automated scores, the effects were stronger when the free writeAlizer R package was employed, whereas for hand-calculated scores, simpler WE-CBM metrics were less costly than more complex metrics. Sensitivity analyses confirmed the relative advantage of automated scores when the number of classrooms, students, and assessment occasions per school year increased; again, writeAlizer was less sensitive to the changes in the ingredients than the other approaches. Finally, the visualization of the cost-effectiveness ratio illustrated that writeAlizer offered the optimal balance between implementation costs and diagnostic accuracy, followed by complex hand-calculated metrics and a proprietary automated program. Implications for the use of hand-calculated and automated scores for the universal screening of written expression with elementary students are discussed.

## 1. Introduction

Considerable evidence suggests that elementary students do not always develop writing skills to levels of proficiency necessary for success in school and beyond (National Center for Education Statistics [NCES], 2012). Effectively identifying students at risk for poor writing performance requires screening measures that are predictive of relevant outcomes, such as scores on statewide achievement tests. Several factors influence the adoption of screening measures in applied settings, such as their technical adequacy, usability, and implementation costs. Although researchers have accumulated evidence related to technical adequacy (McMaster & Espin, 2007; Romig et al., 2017, 2020) and usability (Payan et al., 2019), the investigation of the economic implications of different procedures of writing screening has been ignored. This lack of knowledge can undermine effective allocation of resources and lead to the adoption of

\* Corresponding author at: University of Houston, 403 Farish Hall, 3657 Cullen Blvd., Houston, TX 77204, USA.  
E-mail address: [mmatta@uh.edu](mailto:mmatta@uh.edu) (M. Matta).

suboptimal measures in schools. Therefore, the purpose of this study was to analyze the costs and cost-effectiveness associated with different quantitative scoring approaches for curriculum-based measures (CBM) of writing.

### 1.1. Theoretical model of writing development: the not-so-simple view of writing

According to the Not-so-Simple View of Writing (NSVW) model, effective writing is a function of three sets of cognitive processes: (a) transcription skills for the generation of written text (e.g., handwriting and spelling), (b) text generation skills for the translation of ideas into units of written language (e.g., text structure and genre), and (c) self-regulatory skills for the use of goal-oriented control processes (e.g., planning and goal setting). In addition, efficient working memory allows written expression to be minimally constrained by lower-level skills (e.g., transcription skills) and stimulates a more effective translation of ideas into higher-level units of language (i.e., text generation and self-regulatory skills; Berninger & Winn, 2006).

From a developmental perspective, student written expression initially is highly dependent on transcription skills (Graham et al., 1997). As students develop more efficient lower-level writing processes, text generation skills are likely to become more critical to produce high quality writing samples. These processes interact with multiple levels of written language. At the word level, compositions typically show improved lexical diversity as a function of student wider expressive oral language (Olinghouse & Graham, 2009). At the sentence level, writing samples present more complex syntactic structures (Beers & Nagy, 2011). At the discourse level, upper elementary students learn to organize their writing in relation to specific genres (Galloway & Uccelli, 2015). When developing screening methods for written expression, it is important to capture the complexity at multiple levels of language, especially for upper elementary and older students whose skills have likely moved beyond transcription. The measurement of higher-level text characteristics likely improves the face validity of writing assessments and provides a more accurate identification of students at risk for writing difficulties, particularly at the upper elementary grades.

### 1.2. Hand-scored approaches for the screening of writing in elementary grades

Written Expression CBM (WE-CBM) is a brief, technically adequate, and repeatable measure for the screening of writing fluency (Deno et al., 1980). WE-CBM tasks involve providing a prompt to students who are given 1 min to think about the topic and a brief amount of time (e.g., 3–15 min) to write a response. The written response is then scored for simple countable linguistic indicators, such as total words written (TWW), words spelled correctly (WSC), and correct word sequences (CWS). These indicators are production-dependent in that they are a function of the amount of text generated by the student within the allotted time. Subsequent to the development of these measures, correct minus incorrect word sequences (CIWS) was introduced as a way to account for composition length and capture both writing accuracy and fluency (Espin et al., 2000). When these metrics are used for screening, they are useful for the identification of students who do not perform adequately and are at risk for poor outcomes. In other words, the scores indicate the potential for a problem and the need for more instructional or intervention supports (Hosp et al., 2016; Shapiro, 2010).

The selection of quantitative metrics to use with students often involves a trade-off between feasibility, alignment with the components of written expression, and technical adequacy. Simple metrics that capture surface linguistic features (e.g., total words written) are easy to calculate, relatively inexpensive, highly reliable, require minimal training, and work well as predictors of younger students' overall performance on writing tasks (Jewell & Malecki, 2005); however, their technical adequacy weakens when used with upper elementary and older students, and teachers are generally unwilling to use them because of low face validity (Ritchey & Coker, 2013; Yell et al., 1992) and might be reflective of lower order skills, such as transcription (Allen et al., 2020). Conversely, more complex metrics that capture sophisticated linguistic features or deep textual structure are time-consuming to score and more sensitive to the rater's subjectivity. These metrics are, however, reflective of higher order skills, such as text generation (Allen et al., 2020), and have shown better diagnostic accuracy and predictive validity (Keller-Margulis et al., 2021).

Ultimately, selection of the best approach for screening writing should involve consideration of technical adequacy and usability as well as implementation costs. Identification of the costs associated with all the resources involved in using a measure or intervention is known as *cost analysis* (Institute of Education Sciences [IES], 2020; Levin et al., 2018). Cost analysis results can be used to inform further economic considerations, such as comparing different screening methods, and provide a variety of metrics to determine which approach might be preferable to implement given the available resources.

### 1.3. Feasibility issues of hand-scored approaches to screening writing

Although hand-calculated metrics have evidence of technical adequacy, with CIWS often outperforming the others (McMaster & Espin, 2007; Romig et al., 2017), the implementation of WE-CBM metrics has been associated with low scoring feasibility. Two studies reported the scoring time for 3-min writing samples completed by students in Grades 3 and 4 (Gansle et al., 2002, 2004); on average, trained raters scored the writing samples for TWW in 24.9–25.9 s and for CWS in 57.3–72.1 s. WSC was calculated only in one study and required about 28 s. Notably, Gansle and colleagues did not report data for CIWS, and hence we do not know with certainty if CWS and CIWS would require the same scoring time. Although raters are likely similar in the time required to establish whether a sequence between two words is correct or incorrect, additional time is expected to be necessary to count the total number of incorrect word sequences (IWS) and to subtract IWS from CWS. Based on the available literature, it is reasonable to estimate an average scoring time of over 2 min to score one writing sample for these four WE-CBM metrics. Unfortunately, the scoring time likely increases when class-wide screening procedures are used, when students complete multiple writing tasks, when they have more than 3 min to write, or when older students are screened (Espin et al., 1999).

Given that the use of traditional WE-CBM approaches is time consuming, particularly for those metrics that offer more adequate validity coefficients (i.e., CWS and CIWS), many practitioners determine that it is not a reasonable use of time and resources. We believe that this might partially explain the reasons behind the limited use of WE-CBM by school psychologists (16.3%) as part of universal screening process as compared to CBM for other academic outcomes, such as reading comprehension (29.5%) and math computation (27.7%; Benson et al., 2019). The challenges associated with traditional WE-CBM as well as the ongoing need to effectively measure student performance in the area of writing has led to the exploration of automated text evaluation programs (or automated essay scoring) as alternative options for measuring and making decisions about student writing performance for the purpose of screening (Deane, 2013).

#### 1.4. Automated approaches to screening writing

Although the use of automated programs to score WE-CBM samples had initially little success (Gansle et al., 2002), recently the availability of more sophisticated technology has generated renewed interest. Computer programs can be used to reproduce hand-calculated WE-CBM metrics (Mercer et al., 2021); however, the generation of writing quality scores is likely more consistent with the interpretation and use of WE-CBM scores as brief indicators of general writing proficiency in research and applied settings (see Espin & Deno, 2016, discussing the same issue for reading CBM). Two recent meta-analyses on the criterion validity of WE-CBM scores have shown that fluency-based measures are not commonly used to draw inferences on student writing fluency per se but rather on writing performance on state-developed or commercially developed tests which are typically designed to assess broad writing performance (Romig et al., 2017, 2020). In other words, WE-CBM scores are interpreted as global measures of writing proficiency reflecting both fluency (given the time constraint of WE-CBM tasks) and quality (given the criterion measures used to test the validity of WE-CBM scores).

Researchers can employ computer programs to generate and combine hundreds of linguistic metrics into one or more composite scores of writing quality with greater reliability, similar psychometric properties to hand-calculated systems, and improved feasibility (Keller-Margulis et al., 2021; Matta et al., 2022). The wider range of metrics may also better capture more complex aspects of text generation. Of the automated approaches available, Project Essay Grade (PEG) and writeAlizer have shown promising results for the screening of upper elementary students with writing difficulties. In general, the two programs apply scoring models developed via machine learning techniques to reproduce human judgments of writing quality (Mercer et al., 2019; Wilson et al., 2017). They employ similar technologies to generate automated writing scores and can accurately identify students at risk for poor writing outcomes (Keller-Margulis et al., 2021; Wilson & Rodrigues, 2020). However, they differ in terms of scoring model customization, transparency, and costs.

##### 1.4.1. Project Essay Grade

Project Essay Grade (PEG; Page, 1966, 2003) was the first automated system commercially available for the assessment of overall writing quality along with six specific traits (i.e., conventions, ideas, organization, sentence structure, style, and word choice). PEG scores student writing through *genre-specific* and *prompt-independent* scoring models (C. Palermo, personal communication, September 19, 2019). The system yields hundreds of linguistic metrics for each writing sample and combines them into global or local indicators of writing quality depending upon the writing sample genre (e.g., narrative, expository, persuasive). Alternatively, scoring models can be *prompt-dependent* when a large corpus of essays written from the same prompt are available.

In two studies, PEG classification accuracy was examined in the context of low-stakes decisions for upper elementary students. Wilson (2018) used PEG to score one, 30-min argumentative essay for overall writing quality completed by students in Grades 3 and 4 in fall and spring. Receiver operator characteristic (ROC) curve analysis was used to examine the extent to which PEG scores aligned with proficiency levels on a standardized literacy test and establish cut-off scores for screening to identify those students at risk for poor performance. Consistent with human ratings on analytic rubrics (Lai et al., 2015), the six traits calculated through PEG showed a unidimensional underlying structure, hence only the overall score of writing quality was used for the analysis (Wilson, 2019). Findings showed that diagnostic indices ranged from acceptable to excellent and were consistent across the two time-points ( $AUC_{fall} = 0.74\text{--}0.79$  and  $AUC_{spring} = 0.75\text{--}0.83$ ). Wilson and Rodrigues (2020) then replicated the analyses aiming to generalize the results to older students and other genres of writing. Students in Grades 3–5 completed six, 30-min writing tasks, consisting of two essays per genre, in fall and a standardized literacy test in spring. The diagnostic accuracy of PEG scores was again evaluated via ROC curve analysis and compared with word count. Results indicated that one writing sample was sufficient for the accurate classification of at-risk students in Grades 3 and 4, whereas three samples were needed for students in Grade 5. Moreover, PEG overall quality scores consistently led to more accurate classification than word count, although the two variables were highly correlated ( $r = 0.79\text{--}0.90$ ). However, because PEG is a proprietary system, the scoring models used to weight text characteristics and generate composite scores of writing quality are not publicly available. The lack of such information is a critical limitation given that automated text evaluation models are frequently criticized for disproportionately weighting composition length (Perelman, 2014) and concealing differences across grades and writing genres. For instance, it is reasonable to expect the role of word count in predicting overall quality to decrease as students get older and other aspects of written text, such as aspects of paragraph structure, become more relevant, denoting a better organization of ideas in the text, especially in expository essays.

##### 1.4.2. writeAlizer

More recently, in response to the inherent limitations associated with proprietary programs, the writeAlizer R package was developed as a free, open-source option for the automated evaluation of writing of elementary school students (Mercer, 2020).

Currently, writeAlizer accepts outputs from two free automated programs, namely Coh-Metrix (McNamara et al., 2014) and ReaderBench (Dascălu, 2014) that are designed to analyze cohesion and textual complexity of written compositions. writeAlizer combines Coh-Metrix and ReaderBench outputs into writing quality composite scores. In particular, writeAlizer uses scoring models developed from a machine learning process that contain the coefficients to weight each textual feature assessed by Coh-Metrix or ReaderBench to generate composite writing quality scores. These models were trained using an independent set of timed writing samples to predict human scores of overall quality defined by idea development and idea organization. In other words, writeAlizer scores are indicators of the degree to which the writing samples contain detailed and interesting ideas that are well-organized.

Initial investigations support the use of writeAlizer in the context of universal screening of elementary students. In Mercer et al. (2019), overall writing composite scores were calculated by weighting Coh-Metrix indices and their validity was compared to hand-scored WE-CBM metrics. Students in Grades 2–5 completed one, 7-min WE-CBM narrative task in fall and winter; then, writing samples were hand-scored for WE-CBM metrics and processed through writeAlizer with Coh-Metrix to generate overall writing quality scores in combination with an applied predictive modeling approach. Results indicated that automated scores offered similar levels of structural and external validity to WE-CBM metrics. In Keller-Margulis et al. (2021), 140 students in Grade 4 completed one, 3-min WE-CBM task at three time points (fall, winter, spring) during the school year.<sup>1</sup> Writing samples were hand-scored for four common WE-CBM metrics as well as processed through writeAlizer and PEG. The average scores across three time points were calculated for both hand-calculated and automated metrics and used for the subsequent analyses. Regression models were estimated to examine the degree to which hand-calculated and automated scores would predict student performance on the statewide writing test completed at the end of the year. The variance explained by composite scores generated through writeAlizer with ReaderBench or Coh-Metrix ( $R^2 = 0.29$  and  $0.30$ , respectively) was higher than simple WE-CBM metrics (TWW =  $0.10$  and WSC =  $0.13$ ) and comparable to more complex WE-CBM metrics (CWS =  $0.31$  and CIWS =  $0.35$ ). Additionally, writeAlizer scores and complex WE-CBM metrics showed better predictive validity than PEG ( $R^2 = 0.24$ ), however, the difference was not statistically significant.

Although these new automated approaches to scoring have shown good technical adequacy and improved feasibility over hand-scored approaches, these features alone are not the only properties to consider for the selection of optimal screening measures. In fact, school administrators must also take into account the costs associated with their implementation through economic evaluation.

### 1.5. Economic evaluations for the adoption of screening measures

Economic evaluations are intended to capture the range of investments required to adopt a screening measure and are based on the idea that diagnostic accuracy alone is not sufficient for making decisions about what to adopt and implement in schools. Although common in other fields (e.g., medicine), this area of study has emerged in the education literature in recent years as a way to optimize the use of available, yet finite, resources in light of the benefit that their use provides.

Several types of economic evaluation approaches exist, and they are used depending upon the type of information desired (Levin et al., 2018). Cost analysis and cost-effectiveness analysis, specifically, have been the primary targets of initial investigations in the field of education (e.g., Barrett, Truckenmiller, & Eckert, 2020; Barrett & VanDerHeyden, 2020; Hollands et al., 2015). Cost analysis is the examination of the cost of resources required to implement a screening measure in applied settings. The approach to cost analysis that requires identification and estimates of the value of all the resources (or ingredients) is known as the “Ingredients Method” (Levin et al., 2018). The results from this type of analysis help administrators determine whether a school or district can afford certain screening measures, or cheaper alternatives might need to be considered. If the measure or innovation is too expensive, then consideration of that option typically is ceased and there is no need for further examination of costs. It is important to note that a cost analysis does not include any information regarding the relative technical adequacy or diagnostic accuracy of a particular tool or innovation, it is simply the overall cost of adoption. Cost-effectiveness analysis, does, however, include the elements of basic cost analysis and allows for the examination of the cost of screening measures relative to the degree to which they accurately identify students at risk for poor performance or outcomes. This type of analysis is important because measures may be effective in terms of diagnostic accuracy but unreasonable in terms of the costs required for implementation.

Examining the cost of the various resources, both materials and personnel, required for implementation of an innovation provides concrete evidence regarding the investment of time and money required to engage in a certain activity. Increasingly, federal funding agencies are requiring inclusion of economic evaluations in proposals such that the information generated from the research can inform not only the effectiveness of interventions but also the investment required to achieve that outcome (Schneider, 2020).

### 1.6. Cost analysis and cost-effectiveness of writing measures

To our knowledge, Barrett, Truckenmiller, and Eckert (2020) are the only authors who have examined the costs of various approaches to writing assessment in the context of a brief intervention program. Their study focused on implementation costs and cost-effectiveness of a brief intervention program that uses performance feedback during writing instruction for elementary students. The program was designed as a classwide approach to improve writing fluency and has been found to improve the number of TWW and CWS when implemented regularly (Truckenmiller et al., 2014).

Using the ingredients framework, the authors identified all the resources required for intervention delivery and estimated their

<sup>1</sup> Results from this study were used to derive the cost-effectiveness of different scoring approaches in the current study.

costs through the online platform *CostOut* (Hollands et al., 2015). Specifically, they gathered costs associated with training, preparation of the necessary materials, implementation, and integrity monitoring. Due to the nature of the intervention, there were no facilities costs. A sensitivity analysis was used to calculate the overall costs of the intervention across three scenarios: (a) the original randomized controlled trial (RCT) with research assistants carrying out every step of the intervention, (b) the participation of teachers for materials preparation and to provide performance feedback, and (c) the participation of teacher assistants for materials preparation and teachers for the intervention delivery.

Results indicated that the original RCT was less costly than the scenario where teachers prepared materials and implemented the intervention and slightly more costly than the scenario where teacher assistants prepare materials and teachers implement the intervention. Among the various ingredients, the differences in cost of the scoring procedures for TWW are notable. In the original RCT, one research assistant hand-scored one, 3-min writing sample of the 46 students in Grade 3 requiring 1 h of time per week. Throughout the intervention, the total time for the ingredient of hand scoring by a research assistant was 9 h and the total cost was \$338.67. Sensitivity analysis revealed that the cost would increase to \$597.78 if samples were scored by teachers and would decrease to \$278.19 if scored by teacher assistants. Overall, the results illustrate the difference in costs associated with varying approaches to training and implementation.

There are various activities in schools that require the investment of resources, including the assessment approaches used to make decisions about student writing performance. Schools invest significant resources in terms of both time and materials to implement measurement approaches to track student performance. Any type of data-based decision making requires time and effort to collect. Unfortunately, economic evaluations for the use of screening measures have received considerably less attention than intervention programs. The current study was designed to fill this gap in the literature by comparing hand-scored and automated approaches to scoring WE-CBM tasks for the purpose of conducting universal screening.

### 1.7. Research questions

The use of automated text evaluation programs is a promising approach to scoring student writing samples for the purpose of universal screening; however, no study has examined whether the use of automated programs for the scoring of WE-CBM samples improves cost feasibility of scoring as compared to hand-scored WE-CBM metrics. Ideally, the implementation of a new measurement approach should be predicated upon empirical evidence supporting equal or better cost-effectiveness in comparison to other approaches employed for similar purposes. For this reason, we used secondary data analyses from Keller-Margulis et al. (2021) and Matta et al. (2022) to expand the literature on economic evaluations in education to include screening tools of writing. First, we conducted parallel cost analyses to estimate the implementation costs of hand-scored and automated approaches to score WE-CBM tasks. Second, we conducted a set of sensitivity analyses to examine the extent to which the addition of units of one or more ingredients (e.g., number of classrooms or number of assessments in the school year) would change implementation costs. Third, we compared the cost-effectiveness of hand-calculated and automated approaches by re-calculating the total costs, as well as costs per classroom and per student and matching such data with corresponding technical adequacy data.

## 2. Method

The study was conducted retrospectively using the ingredients method framework (Levin et al., 2018). We identified all the ingredients needed for the use of both hand-scored and automated metrics for the screening of writing skills. Then, we used the website *Cost-out® - the CBCSE Cost Tool Kit* (Hollands et al., 2015) to retrieve cost data for each ingredient. The website is a free tool designed and maintained by the Center for Benefit-Cost Studies of Education (CBCSE) to evaluate and compare the costs of alternative educational tools and programs. In this study, cost analysis and cost effectiveness were conducted with the values for each ingredient updated to the most recent available national averages to promote the generalizability of results. To overcome the inherent limitations of a “static” cost analysis conducted retrospectively, readers may use the interactive spreadsheet made available online (<https://osf.io/82jbg/>) to modify one or more ingredients and examine the changes in the output costs. This will provide decision-makers with dynamic information resulting in various costs across the academic year and increase the potential utility of our analyses for applied practice.

### 2.1. Ingredients for cost analysis

The following sections describe the characteristics of study and the corresponding ingredients used in the cost analysis of the different scoring approaches to WE-CBM. The ingredients include fixed costs (such as personnel training) and ongoing costs (such as task administration and score generation, procedural integrity evaluation, and test materials; see Table 1). We also explain the reasons for not considering facilities and opportunity costs in the analysis.

#### 2.1.1. Training

Training for the study was conducted for administration of the WE-CBM tasks as well as for the scoring of samples. A total of 44 elementary school teachers from two campuses located in the southwest United States participated in the original study data collection (Keller-Margulis et al., 2021; Matta et al., 2022); they received a packet of materials with detailed instructions for WE-CBM task administration and a checklist with the procedures to follow. Teachers were required to review the materials and contact the researcher or a designated colleague with questions.



**Table 1**

Characteristics of the study, costs for ingredients, and total and marginal costs for each scoring approach.

Characteristics of the study	Number		
Students/Average of students per classroom	722/16.41		
Teachers/Participating classrooms	44/44		
Duration of task completion	3 min		
Writing task per time point	1		
Time points over school year	3		

Ingredients	Hand-scored WE-CBM	writeAlizer	PEG
<b>Training</b>			
Trainer	\$351.20	\$87.80	\$87.80
Research assistants	\$365.44	\$91.36	\$91.36
Teachers	\$2695.44	\$2695.44	\$2695.44
<b>Implementation</b>			
Test administration	\$943.40	\$943.40	\$943.40
Transcription		\$879.49	\$879.49
Cross-check transcriptions		\$467.23	\$467.23
Creation of text files		\$274.84	
Scoring of writing sample	\$3861.52		
Total word written (TWW)	\$590.91		
Word spelled correctly (WSC)	\$989.43		
Correct word sequences (CWS)	\$1937.63		
Correct minus incorrect word sequences (CIWS)	\$2281.18		
writeAlizer (with Coh-Metrix or ReaderBench)		\$460.95	
PEG			\$2166.00
Reliability check	\$772.30		
<b>Materials</b>			
Photocopies	\$281.58	\$281.58	\$281.58
Total cost	\$9270.89	\$6182.10	\$7330.73
Cost per classroom	\$210.70	\$140.50	\$166.61
Cost per student	\$12.84	\$8.56	\$10.15
Marginal cost for 1 additional classroom	\$194.41	\$125.95	\$168.94
Marginal cost for 1 additional student	\$6.81	\$2.64	\$5.26
Marginal cost for 1 additional time-point per classroom	\$44.38	\$25.06	\$35.89

A researcher with expertise in WE-CBM trained four graduate research assistants enrolled in a doctoral-level school psychology program. Although research assistants were at different stages of their doctoral program, they all had received some training in WE-CBM prior to participating in the study (e.g., psychoeducational assessment courses). The main goal of our 4-h training was to provide students with homogeneous scoring guidelines and to ensure adequate interrater reliability. The training was structured in four 60-min sessions. The first session included (a) an overview of the role of fluency for the prediction of student writing outcomes, (b) standardized directions for the administration of WE-CBM tasks, (c) scoring procedures of the four main WE-CBM metrics, and (d) guidelines for the transcription of writing samples from paper to electronic format. The other three sessions involved research assistants scoring writing samples previously collected for WE-CBM metrics. Research assistants individually scored writing samples; then, they compared scores in a group setting and settled disagreements by reviewing the guidelines in the AIMSweb technical manual (Powell-Smith & Shinn, 2004). Additional writing samples were assigned to research assistants between training sessions and scores were reviewed at the beginning of each meeting. During the practice sessions, the trainer calculated the agreement between the established scoring keys and the scores of each research assistant to ensure adequate reliability levels. Research assistants were required to score 20 writing samples and obtain an agreement above 90% with the scoring key. Only after reaching this criterion, they started scoring the WE-CBM samples for the study. Those who failed to do so attended one more training session in which they reviewed the scoring key and were asked to score an additional 20 samples. No research assistant needed more than one additional training session to reach scoring proficiency.

Training costs include hourly salary plus benefits for those implementing the writing assessment (e.g., trainer, classroom teachers, research assistants). The *CostOut* website reports personnel costs based on 10 different sources covering a wide variety of professions and consecutive years adjusted for inflation. The most recent available data indicated that wages and benefits for personnel were as follows: trainer = \$87.80/h; teachers = \$61.26/h; and research assistants = \$22.84/h.

Overall, training costs differed across hand-calculated and automated scoring approaches. Traditional WE-CBM required 4 h for the trainer (4 h x \$87.80/h) for a total of \$351.20, 4 h for each of the four research assistants (4 h x 4 x \$22.84/h) for a total of \$365.44, and 1 h for each of the 44 elementary teachers (1 h x 44 x \$61.26/h) for a total of \$2695.44. Conversely, training costs for the use of automated programs required 1 h for the trainer as well as for each research assistant. This resulted in 1 h of work for the trainer (1 h x \$87.80/h) and 1 h for each of the four research assistants (1 h x 4 x \$22.84/h) for a total of \$91.36. Time and salary for each of the 44 elementary teachers were the same as the training for the hand-calculated scoring approach.

**Table 2**

WE-CBM prompts by grade and time.

Grade	Time	Prompt
2	Fall	I once had a magic pencil and...
2	Winter	He crossed his fingers and opened the box. Suddenly...
2	Spring	The noise was getting louder and louder...
3	Fall	The children were practicing for the school play and...
3	Winter	My 2-year-old brother found a magic marker and...
3	Spring	I was fishing in the river when I felt a terrific tug on the line and...
4	Fall	Yesterday, a monkey climbed through the window at school and...
4	Winter	The bus driver had a bus full of children when it drove into the mysterious fog...
4	Spring	The two space invaders stepped out of their spaceship and...
5	Fall	If I could trade places with my teacher, I would...
5	Winter	Working madly in my science lab, I suddenly realized that my magic formula...
5	Spring	When you are walking down the street one day, a limousine pulls up beside you. When the person inside rolls down the window, you realize that it is the President of the United States. Tell what happens next.

### 2.1.2. Screening task administration

Teachers administered one, 3-min writing task to 722 students in Grades 2–5 at three time-points (i.e., fall, winter, and spring) during the school year. On average, approximately 16 students from each classroom participated in the study. Narrative writing prompts were drawn from the aimsweb system of CBM ([www.aimsweb.com](http://www.aimsweb.com)) and were different across grade and time points (Table 2). Prompts were deemed grade appropriate using professional judgment. Where needed, a word was changed to ensure appropriate vocabulary for the particular grade level. For example, the Grade 3 fall prompt originally included the word “rehearsing” but was changed to “practicing” because that was deemed more grade appropriate. Task administration required approximately 7 min in total, including 3 min to explain directions and answer questions, 1 min of planning, and 3 min for students to write a brief story. Upon the completion of the task, teachers collected the writing samples and provided them to the researchers along with an implementation checklist. Screening task administrations were the same regardless of the scoring approach, resulting in 7 min of work for each of the 44 teachers who participated in the study (7 min × 44 × \$61.26/h) for a total of \$943.40.

### 2.1.3. Procedural integrity

During the administration of the writing tasks, teachers were asked to complete the Accuracy of Implementation Rating Scale WE-CBM (AIRS-WE-CBM; Powell-Smith & Shinn, 2004), a checklist composed of 15 items describing the steps to follow for accurate and reliable administration of the writing prompts. In addition, the administration sessions were audio-recorded at the spring time-point and reviewed by research assistants. Results indicated 100% accuracy.

### 2.1.4. Generation of test scores

**2.1.4.1. Hand-Scored Approaches.** Upon receiving materials from teachers, the four research assistants hand-scored the writing samples for the WE-CBM metrics. Table 3 includes the average scores and the standard deviations across the three WE-CBM tasks by grade as well as the performance of students in Grade 4 on the state writing test.

Although research assistants hand-scored the writing samples for WE-CBM metrics, a postdoctoral fellow captured the scoring time based on a subset of randomly selected samples written by students in Grade 4 ( $n = 20$ ). Research assistants were instructed not to use shortcuts for the calculation of different metrics for each sample. For example, raters were asked not to count the number of words spelled incorrectly while measuring the number of total words written in order to be able to isolate scoring time for each metric. For the calculation of scoring time, we used samples written by students in Grade 4 because (a) their performance represented a good approximation of the average of WE-CBM scores across the four grade levels, and (b) the results from the cost analysis could be used directly for the calculation of cost-effectiveness given that the state writing test was not given to students in other grade levels. The average scoring time was 43 s for TWW, 1 min 12 s for WSC, 2 min 21 s for CWS, and 2 min 26 s for CIWS (see Table 4). On average, research assistants scored one writing sample in 4 min 41 s; the total time was obtained by summing TWW, WSC, and CIWS given that

**Table 3**

Hand-calculated WE-CBM scores by grade.

Metrics	Grade 2		Grade 3		Grade 4		Grade 5	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
TWW	25.89	12.14	34.65	11.85	39.65	10.06	42.71	12.08
WSC	22.51	11.40	31.71	11.63	36.97	9.63	40.26	11.56
CWS	18.15	10.77	27.23	11.88	30.85	9.45	37.32	12.31
CIWS	7.36	11.05	16.32	13.54	18.11	11.88	28.04	14.53
STAAR <sup>1</sup>					26.51	5.74		

Note. Sample size: Grade 2,  $n = 200$ ; Grade 3,  $n = 161$ ; Grade 4,  $n = 181$ ; Grade 5,  $n = 180$ .

<sup>1</sup> 23% of the students did not meet the grade-level expectations on the state writing test.

the scoring time for CWS was already accounted for the calculation of CIWS. This resulted in a total of 2166 samples scored in 169 h 4 min 6 s for a total cost of \$3861.52.

Additionally, two research assistants randomly selected and scored 20% of WE-CBM writing samples ( $n = 433$ ) across all time points and grades. Concordance correlation coefficients ( $\rho_c$ ) were 0.99 for TWW, 0.99 for WSC, 0.96 for CWS, and 0.85 for CIWS and served as evidence of good to excellent interobserver agreement (IOA). Total time to complete the WE-CBM reliability check was 33 h 47 min 53 s for a total of \$772.30.

**2.1.4.2. Automated Approaches.** Research assistants then transcribed the 2166 writing samples from paper to electronic format. The writing samples were transcribed ensuring maximum fidelity to the originals by including errors and hard returns such that samples were typed exactly as written. While research assistants completed the task, a postdoctoral fellow measured transcription time on a subset of randomly extracted samples ( $n = 40$ ) written by students in Grade 4. On average, research assistants transcribed one sample in 64 s. The total transcription time for 2166 samples was 38 h 30 min 24 s with a total cost of \$879.49. All transcriptions were double-checked for accuracy; discrepancies between research assistants were mainly due to students' poor handwriting legibility and accounted for less than 1% of the words across all the writing samples. Time to review one sample for accuracy required on average 34 s, hence a total of 20 h 27 min 24 s and a total cost of \$467.23.

**2.1.4.2.1. writeAlizer.** Graduate assistants saved each of the samples into separate text files in order to create a suitable format for digitized samples to be processed through writeAlizer (Mercer, 2020). On average, the creation of one text file took 20 s for a total of 12 h 2 min and a total cost of \$274.84. Then, a researcher with expertise in automated text evaluation, psychometrics, and proficient in the use of the R software (RStudio Team, 2020) processed the writing samples through automated programs and generated two sets of writeAlizer composite scores: (a) writeAlizer with Coh-Metrix, and (b) writeAlizer with ReaderBench.<sup>2</sup> In particular, writeAlizer imports ReaderBench and Coh-Metrix output files into R and uses scoring models to weight and combine text features into one composite score of predicted writing quality. The scoring models containing information about relative importance of the text features (i.e., beta coefficients) were developed from independent 7-min WE-CBM samples. More information on the scoring models is available on the writeAlizer GitHub website. Salary for the data analyst was \$87.80/h. The generation of automated scores required 1 h 45 min at each of the three time points WE-CBM tasks were administered. This resulted in a total time of 5 h 15 min and a total cost of \$460.95 over the course of one school year.

**2.1.4.2.2. PEG.** Consistent with the purpose of the study, WE-CBM samples were processed through PEG by Measurement Incorporated (Page, 2003) which used a narrative prompt-independent scoring model to calculate writing quality scores on six traits. Considering the unidimensionality of the six traits, we calculated the sum of the six traits and used it as an indicator of overall writing quality for data analyses. A flat rate of \$1 was charged to process each of the 2166 writing samples. Therefore, the total cost to generate PEG scores was \$2166. Although PEG might be used in this way, it is more common that students would type their own writing samples in the online platform. Moreover, students have access to the platform over the school year and can instantly receive scores and feedback on their written production. Of course, this has implications for the total costs and possibly the diagnostic accuracy over time. However, in the current study, we included the flat rate instead of the cost per user for two reasons: (a) the comparison among scoring approaches focused on writing screening, and (b) the final cost of PEG would be unfairly inflated by the fees associated to the individual accounts on the platform.

## 2.1.5. Materials

Teachers were provided with a packet for their classroom with a response sheet for each student that included the story prompt and enough space to write for 3 min. The materials were estimated at \$0.39 per student over the school year (approximately \$0.13 per paper  $\times$  3 units/school year) for a total of \$281.58. CostOut retrieved information about 2020 price of pages printed in black and white from staples.com.

## 2.1.6. Facilities

All activities took place in classrooms during school hours. Space and electricity were not considered as ingredients in the current analysis because their inclusion would lead to overestimation of implementation costs (Crowley et al., 2018). In addition, their use was not tied to the specific task at hand (i.e., students would be occupying space with the lights on regardless because data were collected during the school day).

## 2.2. Diagnostic accuracy of the scoring approaches

The cost-effectiveness of the approaches to scoring were compared using research data on their diagnostic accuracy. This approach was devised to serve as a parallel to the incremental cost effectiveness ratio often used in cost-effectiveness analysis but for which there is no designated process for examining measurement approaches.

As reported in Keller-Margulis et al. (2021), a study was conducted to examine the extent to which hand-calculated and automated scoring approaches to WE-CBM would accurately identify students at risk for poor writing outcomes. Students in Grade 4 from the

<sup>2</sup> Although we processed writing samples in batches for both programs, currently Coh-Metrix allows to process only one sample at a time. Therefore, the reader interested in using automated approaches to scoring writing should consider ReaderBench as the preferred option.



**Table 4**  
Processing and scoring time.

Process	M
Hand-scored WE-CBM	
Scoring one 3-min writing sample for all metrics	4 min 41 s
Scoring one 3-min writing sample for TWW	43 s
Scoring one 3-min writing sample for WSC	1 min 12 s
Scoring one 3-min writing sample for CWS	2 min 21 s
Scoring one 3-min writing sample for CIWS	2 min 46 s
Automated scores	
Transcription of one 3-min writing sample	1 min 4 s
Cross-check of one 3-min transcription	34 s
Creation of one text file (writeAlizer only)	20 s
Generation of scores from automated programs (writeAlizer only)	1 h and 45 min

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

sample described above completed one, 3-min WE-CBM task at the three time points and took the statewide writing test at the end of the year. Writing samples were then hand-scored with WE-CBM metrics, digitized using a transcription process, and processed through the writeAlizer R package (Mercer, 2020) and PEG to generate composite scores of writing quality. Then, we calculated the average score across the three time points for both hand-calculated (i.e., TWW, WSC, CWS, and CIWS) and automated metrics (writeAlizer with Coh-Metrix, writeAlizer with ReaderBench, and PEG) and estimated diagnostic accuracy coefficients. Receiver operating characteristic (ROC) curves and AUC values were used to estimate the probability that a student who failed the statewide test would be rated as more likely to fail based on the scores of each metric. Results indicated that hand-calculated metrics showed comparable or poorer diagnostic accuracy than automated scores (see Table 5). TWW and WSC (AUC = 0.69 and 0.73, respectively) underperformed predicted quality scores generated through writeAlizer with ReaderBench and Coh-Metrix (AUC = 0.81 and 0.82, respectively) as well as through PEG (AUC = 0.83). More complex WE-CBM metrics, such as CWS and CIWS, yielded AUC values of higher magnitude (AUC = 0.84 and 0.89, respectively), but not significantly different from either the writeAlizer scores or PEG.

### 2.3. Analytic strategy

The present cost analyses were carried out based upon characteristics of the Keller-Margulis et al. (2021) and Matta et al. (2022)'s studies. From there, we derived the number of participants (i.e., trainer, data scientist, research assistants, teachers, and students), tasks per time point, time points per school year, and task duration (as described above) in order to calculate the costs of ingredients for training, task administration, score generation, and materials for the two scoring approaches. Missing data (i.e., students who did not complete writing samples at one or two time points) were ignored for the calculation of the costs in that (a) policymakers interested in implementing WE-CBM tasks generally cannot forecast the number of students who will be absent from school when they allocate the resources in the budget for the school year, and (b) outside the research setting, students unable to complete the assessment can easily take the test on the next available school day, hence reducing the likelihood for educators of dealing with missing data. The opportunity costs for teachers participating in the study (defined as monetary value associated with using the time in other ways) were considered negligible given that teachers were required to review materials for task administration at the beginning of the school year for 1 h and the administration of WE-CBM tasks only took 7 min per time point for a total of 21 min in one year. Additionally, assuming no turnover, the training in WE-CBM received in the context of the study can be used in the future.

Three parallel cost analyses were conducted to estimate the total costs for implementation of each scoring approach. The first

**Table 5**  
Total and marginal costs disaggregated per metric and corresponding AUC values.

Scoring approach	Total cost	Cost per classroom	Cost per student	AUC <sup>a</sup>
Hand-Calculated WE-CBM				
TWW	\$5346.15	\$121.50	\$7.40	0.69
WSC	\$5824.38	\$132.37	\$8.07	0.73
CWS	\$6962.22	\$158.23	\$9.64	0.84
CIWS	\$7374.48	\$167.60	\$10.21	0.89
Automated scores				
wA:CM	\$6182.10	\$140.50	\$8.56	0.82
wA:RB	\$6182.10	\$140.50	\$8.56	0.81
PEG	\$7330.73	\$166.61	\$10.15	0.83

Note. TWW = Total Words Written; WSC = Words Spelled Correctly; CWS = Correct Word Sequences; CIWS = Correct Minus Incorrect Word Sequences; writeAlizer:CM = writeAlizer based on Coh-Metrix scores; writeAlizer:RB = writeAlizer based on ReaderBench scores; PEG = Project Essay Grade.

<sup>a</sup> AUC = Area Under the Curve calculated for the fail/pass criterion on the statewide writing test. The AUC values are derived from Keller-Margulis et al. (2021).

analysis focused on hand-scored WE-CBM total cost and included the costs required to train personnel and hand-score writing samples for four WE-CBM metrics across three time points. The second and third analysis focused on writeAlizer and PEG total cost respectively and included the costs required to train personnel, digitize writing samples, and generate writing quality predicted scores from automated programs. Although PEG is an automated platform for the evaluation of written expression, it was kept separate from the other programs because its use involves different procedures and expenditures. To calculate the costs per classroom and per student, the total costs for the implementation of different approaches were divided by the number of classrooms participating in the study and students completing the writing tasks.

Three sensitivity analyses were then conducted to determine marginal costs associated with the inclusion of one additional classroom, students, or assessments during the school year (e.g., four time points instead of three). Subsequently, total costs as well as costs per classroom and costs per student were calculated for hand-calculated WE-CBM metrics and writeAlizer and PEG composite scores separately. Costs were considered along with diagnostic accuracy data to identify which metrics were more cost-effective.

### 3. Results

Table 1 includes the characteristics of the study as well as total costs for the ingredients involved in the assessment implementation, costs per classroom and per student, and marginal costs associated with changes to some ingredients. Results indicate that the implementation of writeAlizer (\$6182.10) and PEG (\$7330.73) were less expensive than the four hand-calculated WE-CBM metrics (\$9270.89). A similar pattern was found for the total costs per classroom and per student. Costs per classroom were \$140.50 for writeAlizer and \$166.61 for PEG, and \$210.70 for hand-scored WE-CBM, whereas cost per student was \$8.56 for writeAlizer, \$10.15 for PEG, and \$12.84 for hand-scored WE-CBM.

The difference in cost between hand-scored and automated approaches can be attributed to specific ingredients. Training for teachers, time for task administration, and cost for materials were identical because the activities performed by teachers and students were the same across conditions. Higher WE-CBM costs were associated with the time and personnel costs associated with training graduate research assistants, time for hand-scoring of writing samples, and ensuring scoring reliability. By contrast, the ingredients

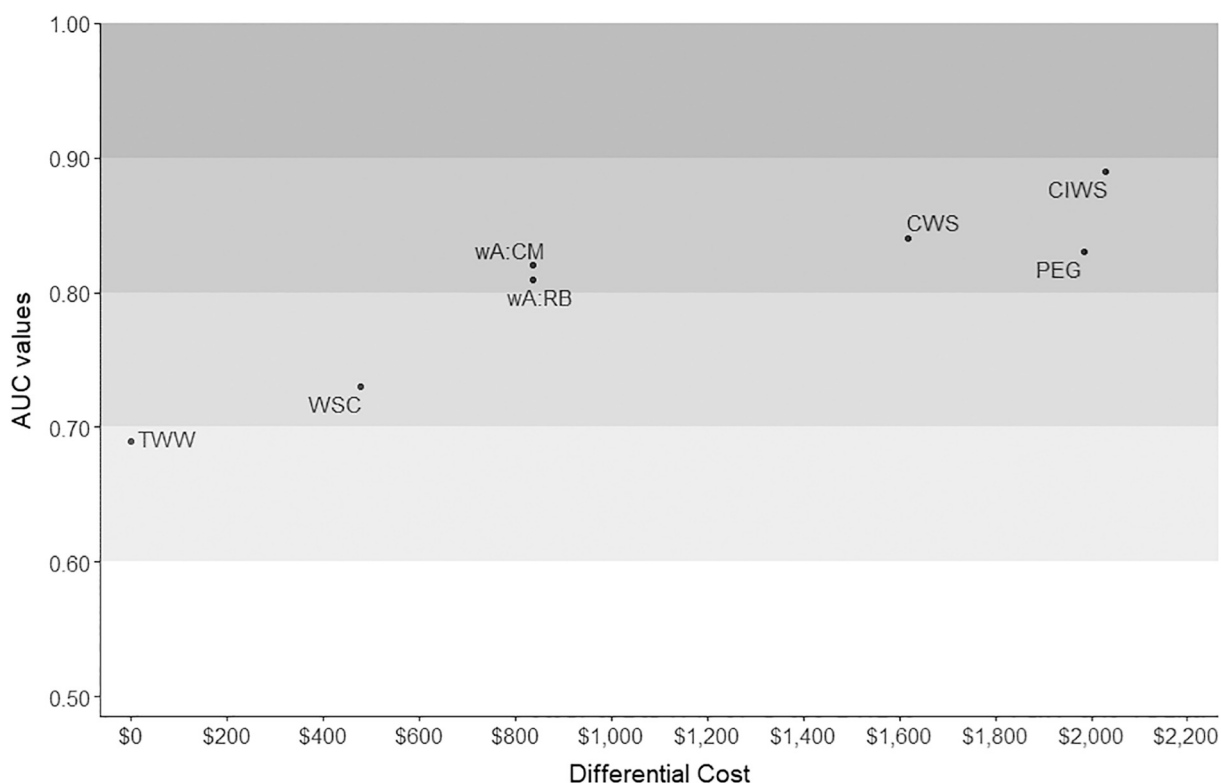


Fig. 1. Ratio of differential costs to AUC values for each scoring approach.

Note. TWV = Total Words Written; WSC = Words Spelled Correctly; CWS = Correct Word Sequences; CIWS = Correct Minus Incorrect Word Sequences; writeAlizer:CM = writeAlizer based on Coh-Metrix scores; writeAlizer:RB = writeAlizer based on ReaderBench scores; PEG = Project Essay Grade.

Differential Cost is expressed as the difference of each metric or scoring approach with TWV. Grey bands indicate different levels of diagnostic accuracy; these can be interpreted as follows: 0.50 = chance; 0.50–0.70 = poor accuracy; 0.70–0.80 = acceptable accuracy; 0.80–0.90 = good accuracy; 0.90–1.00 = excellent accuracy (Hosmer et al., 2013).

that were most costly when using automated programs were the digitization of writing samples and the generation of scores from text evaluation programs.

Results from sensitivity analyses indicate that the addition of one more classroom would be cheaper for writeAlizer (\$125.95) and PEG (\$166.61) than hand-scored WE-CBM (\$194.41). A similar pattern would occur for the addition of one more assessment per classroom in the school year; the marginal costs ranged between \$25.06 for writeAlizer and \$44.38 for hand-scored WE-CBM. Finally, writeAlizer and PEG were also less expensive (\$2.64 and \$5.26, respectively) than hand-scored WE-CBM (\$6.81) for the addition of one student per classroom.

Table 5 includes the costs associated with the use of single hand-scored WE-CBM metrics. TWW and WSC were the least costly metrics (\$5346.15 and \$5824.38, respectively), whereas CWS and CIWS the most expensive (\$6962.22 and \$7374.48, respectively). Patterns were similar for costs per classroom and costs per student. Costs per classroom varied from \$121.50 to \$167.60 and costs per student ranged from \$7.40 to \$10.21 for hand-calculated WE-CBM. The disaggregated costs for individual WE-CBM metrics allow for a more direct comparison with the automated approaches to scoring; simple WE-CBM metrics (such as TWW and WSC) were expected to be the least expensive options to use, whereas writeAlizer was far cheaper than both more complex WE-CBM metrics (i.e., CWS and CIWS) and PEG. However, any comparison based on the total costs is limited in that it does not consider the extent to which the metrics accurately identify at-risk students. Therefore, it is important to interpret the total costs in light of its performance with vulnerable populations.

Cost-effectiveness of WE-CBM was assessed by comparing implementation costs and the AUC values (see Fig. 1). As noted, there is no accepted approach for this type of analysis when examining measures so a visual approach to understanding the ratio between costs and effectiveness was used. TWW and WSC were the least expensive WE-CBM metrics to implement, but also the most ineffective for decision-making in that they demonstrated poor to fair diagnostic accuracy. Conversely, using TWW as the reference, CWS and CIWS were more expensive but offered far better diagnostic accuracy.

The use of writeAlizer scores was more cost-effective than the implementation of hand-scored WE-CBM metrics or PEG. writeAlizer with ReaderBench or Coh-Metrix showed good diagnostic accuracy. CWS, CIWS, and PEG performed within the same range of technical adequacy. However, the differential costs of writeAlizer ( $\Delta\text{cost} = \$835.94$ ) were far less expensive than CWS ( $\Delta\text{cost} = \$1616.07$ ), PEG ( $\Delta\text{cost} = \$1984.57$ ), and CIWS ( $\Delta\text{cost} = \$2028.33$ ).

#### 4. Discussion

The study of cost in education has recently received increased attention in the literature and in practice (Barrett, Gadke, & VanDerHeyden, 2020). The implementation of assessment measures, like many other decisions about products for use in schools, comes with both fixed and variable costs. Decisions regarding their adoption should consider the costs associated with each ingredient (cost analysis), the degree to which costs might change as a function of modifications of assessment procedures (sensitivity analysis), and the costs as they relate to measures of diagnostic accuracy (cost-effectiveness).

The purpose of this study was to illustrate the cost of implementing various approaches to screening for written expression and to demonstrate that automated programs to score WE-CBM samples have the potential to improve scoring feasibility as compared to traditional WE-CBM metrics. We found that the use of free automated programs was less expensive than both traditional scoring systems, which use human raters, and another commercially available automated program for writing samples generated outside its online platform (i.e., PEG). We also illustrated how our findings generalize to scenarios with different implementation procedures, such as more screening sessions during the school year and inclusion of a larger number of students. Finally, we illustrated that free automated programs were cost-effective solutions given their implementation was less expensive than other approaches to scoring while offering similar or improved accuracy for the identification of students at risk for poor writing outcomes.

The results of this study make four unique contributions to the literature regarding economic evaluations for screening methods of writing skills. First, we identified the costs of ingredients used for the implementation of different scoring methods and showed the extent to which they contribute to the total costs of implementing WE-CBM. Overall, the costs for personnel training and scoring procedures account for the largest differences in the total costs between automated and hand-scored, traditional WE-CBM procedures. With regards to personnel training, research assistants require fewer training sessions to score the writing samples reliably. Although training for traditional WE-CBM involves numerous sessions to familiarize raters with scoring guidelines and substantial practice on existing samples to achieve acceptable interrater reliability, the use of automated programs simply requires personnel be capable of typing writing samples and creating text files on a computer. Regarding implementation costs, research assistants generate scores of writing performance in less time; although WE-CBM requires human raters to read and score writing samples sequentially, the application of scoring models to the outputs of automated programs is conducted for all students simultaneously. In other words, the workload to score writing samples for WE-CBM metrics increases linearly with the number of participating students, whereas the time to evaluate student performance via computer-based programs remains the same. Admittedly, when using automated programs, the digitization of the writing samples also increases with the number of students involved in the screening, but the process takes approximately one-third of the time required for human raters to score one sample for the four primary WE-CBM metrics.

Not only can these data guide selection of the instruments to implement in schools, they also identify expensive ingredients within each scoring approach to further inform decisions about use given the resources available. For instance, the costs associated with the ingredients of automated approaches show that approximately 20% of the total budget is spent for transcriptions. It is reasonable that schools with computer labs could schedule screening sessions allowing students to type their stories, which would then be available for further processing (Protopapas & Skaloumbakas, 2007). This change would eliminate the time and salary required for transcription of the samples and reduce the total costs of implementation. The costs for transcriptions would not be calculated for the typical

implementation of PEG where students would type the writing samples directly onto the online platform.

Second, we conducted a series of sensitivity analyses to calculate the extent to which total costs of different methods change as a function of the number of classrooms, students, or screening sessions. Consistent with total costs, an increase in the quantity of ingredients leads to higher costs for WE-CBM as compared to the same changes introduced for automated programs. These higher costs are due to the increase in salary for personnel who will need more time to score a larger number of writing samples. Conversely, the number of classrooms or students participating in the screening process would have minimal impact on the costs associated with the use of automated programs. Certain programs represent far cheaper solutions than others; for example, the use of free software (such as ReaderBench) allows for scoring large batches of writing samples at the same price regardless of the number of classrooms or students involved, whereas proprietary programs (such as PEG) might increase for each new student involved or for the inclusion of new samples to score.

Third, we found that free computer programs were cost-effective options for writing scoring. Cost-effectiveness was assessed for each metric separately because the parameters considered in the ratio varied greatly both between and within each scoring approach. According to the classification criteria suggested by Hosmer et al. (2013), TWW and CWS were the only two metrics to show poor to acceptable diagnostic accuracy. Following the IES guidelines, a cost-effectiveness plan makes it easier for policymakers and researchers to engage with and interpret results (Hollands et al., 2021). A formal comparison of cost-effectiveness ratios among writeAlizer, PEG, and WE-CBM metrics would be possible upon the availability of cut-off scores indicating risk of poor writing performance. Unfortunately, given the early stage of this work, such cut-off scores or benchmarks for performance on writeAlizer are yet to be established. When cut-off scores become available, the number of students correctly identified with this scoring approach can be used as the denominator and the results can be interpreted as the monetary value of the correct identification of one student.

Furthermore, this study showed that automated scores developed through machine learning algorithms trained to reproduce human scores offered a less expensive and more cost-effective approach to scoring writing. Notably, complex WE-CBM scores are also good indicators of writing quality given that there is evidence of strong correlation coefficients between CWS and human ratings of writing quality (e.g., Ritchey & Coker, 2013). This means that the complexity of features assessed through the scoring method (rather than simply whether samples are scored by human raters or computer programs) better capture core lower and higher order skills underlying written language and lead to stronger classification accuracy. Automated scores demonstrate other important advantages, such as allowing for the customization of scoring models by grade and writing genre, enabled by the variety of metrics at multiple levels of language (i.e., word-, sentence-, and discourse-level) that they use to generate the composite scores; in addition, the automated programs eliminate the potential concern of low agreement among different raters in that they rely on objective indices. The inclusion of more sophisticated aspects of written expression makes the use of automated procedures potentially advantageous for formative assessment of students in elementary school and beyond. For example, this approach offers improved feasibility for scoring both short and long compositions and allows for the application of flexible scoring models that weight linguistic metrics differently as a function of a student's grade and writing genre. Automated programs generate scores of writing quality that are not affected by reliability issues, unlike hand-calculated approaches, and the application of validated scoring models guarantees the generation of the same scores for the same text regardless of training, expertise, central tendency effects, and external factors (such as rater fatigue; Leckie & Baird, 2011; Wilson et al., 2017). Moreover, although the students' spelling errors were maintained in the transcription process, hence preventing those words from being matched to the word lists in the automated process, writeAlizer scores have shown good validity with human judgments of writing quality on the evaluated samples, even with this potential noise in the underlying Coh-Metrix and ReaderBench scores (Keller-Margulis et al., 2021; Matta et al., 2022).

Ultimately, the findings of this study may inform the selection of cost-effective methods for universal screening of written expression in elementary school. The findings might also be used as a comparison to calculate the costs of other methods and to evaluate whether alternative, less expensive approaches with similar or improved accuracy might be available. If administrators wanted to estimate the costs of implementing a writing screener used in their district, the steps described in this paper and the online interactive spreadsheet could serve as a roadmap. First, administrators would need to identify all the ingredients related to implementation and obtain information about the diagnostic accuracy of the screener. Then, the total costs and the cost-effectiveness might be compared to other approaches, including hand-scored and automated approaches to WE-CBM. However, administrators might want to consult with school psychologists or other qualified professionals to ensure appropriate application of this process (Barrett, Gadke, & VanDerHeyden, 2020). Educational professionals with expertise in formative assessment can facilitate the consideration of long-term implications of the use of different screening measures.

These elements must also be understood in the particular context where they are applied. For example, universal screening is typically linked to other tiers of support in multi-tiered systems of support frameworks, and cost and resource allocation are effectively used when data collected from students inform instruction and intervention delivery to address individual needs. Among the automated approaches to scoring WE-CBM samples, PEG has higher implementation costs and provides similar levels of diagnostic accuracy as compared to other free text evaluation programs. The PEG system, however, not only scores student writing samples for writing quality, but it is also connected to MI Write. The MI Write system is a separate, web-based tool designed as an instructional aide where students can enter writing and instantly receive scores aligned to six traits of effective writing and one composite score of overall quality. Through the system, students also receive feedback about their writing as well as various suggested activities identified to improve performance (Wilson & Roscoe, 2020). Thus, although the cost to purchase PEG is higher when only examining the use of this tool for screening, it is important to note that the cost is associated with the various other functions that the system provides and not just a measure for screening.

The need to accurately and feasibly measure student writing performance in ways that allow for the identification of students at risk for poor performance is driven by conclusive data indicating that students struggle to achieve adequate performance in this basic skill

area (NCES, 2012). As outlined in the present study, existing measurement approaches for scoring WE-CBM samples require significant resources for scoring and the metrics generated may not provide data of sufficient technical adequacy for decision-making. The results of recent research suggest that automated text evaluation or automated essay scoring tools may be a viable alternative to traditional approaches to scoring WE-CBM in terms of the technical adequacy of the scores that are produced as well as the accuracy and efficiency associated with scoring. Improved technical adequacy is necessary but not sufficient in evaluating an alternative option for use in screening to identify students at risk for poor performance.

## 5. Limitations

The results described here should be interpreted in light of several limitations. First, the costs included in this study were calculated using national averages for various resources. Although the use of national averages is a common approach in this type of cost demonstration, the true cost of various resources is likely to differ across the country. Nevertheless, using these numbers allows for comparisons in the literature. In practice, when conducting these evaluations locally to inform the actual decisions of schools and districts, the use of numbers relevant for the specific context is essential.

Second, research assistants scored writing samples for WE-CBM metrics and transcribed the text into a digital format, and a researcher familiar with R generated automated scores. However, this type of personnel might not be available in most schools. In fact, the implementation of hand-calculated and automated approaches in applied settings would likely involve teachers scoring WE-CBM and typing writing samples on a computer and an administrator (possibly at the district-level) for the generation of automated scores. Because teachers receive a higher salary compared to research assistants, this change would increase the total costs of each scoring approach and, given the higher weight of scoring for WE-CBM than transcriptions on the total costs, would further improve the cost-effectiveness of the automated programs. Therefore, our results should be interpreted as conservative estimates of the cost-effectiveness of the use automated programs for the scoring of writing.

Third, it should be noted that the cost estimations used for the WE-CBM were not based on newly collected data and were instead extrapolated from the limited existing literature regarding the time required to accomplish hand scoring of WE-CBM samples. The amount of time required for scoring will be directly related to the duration of time used to generate the samples as well as the grade level of the students providing the samples with older students and longer samples increasing the scoring time required. Interested readers might use the spreadsheet available online to examine the degree to which the number of total words influences the estimated costs. Moreover, only students in Grade 4 contributed to the calculation of the AUC values for the screening approaches. Differences in the cost-effectiveness across the approaches for other grades might be possible. Additional work is necessary to generalize the results of this study to other elementary school grades. Furthermore, the use of the average score from WE-CBM tasks given at three different timepoints during the school year does not represent a practical approach; however, it provided some insight regarding the potential value of basing scores on multiple screening samples. Future research will need to establish the number of samples needed for optimal validity and the extent to which other components of the assessment (e.g., task duration, writing genre, scoring approach) can change total costs and diagnostic accuracy estimates. The characteristics of the task might affect the nature of the construct. For example, when the writing task requires students to compose text for a short (e.g., 3 min) vs long period of time (e.g., 60 min), the construct underlying the assessment might differ even if the same scoring approach is used. Of the different scoring approaches, WE-CBM scores might depend more on the characteristics of the task with shorter durations engaging students in lower-level processes of writing abilities (e.g., transcription) vs. longer durations in higher-level processes (e.g., self-regulation and cohesion). Because we intended to compare the costs between hand-calculated and automated approaches for WE-CBM samples, this aspect of the study likely affected the approaches similarly and would not change the ultimate conclusions.

A fourth limitation to note is that we did not account for any of the time or resources required for school personnel to review and make decisions about the data collected. Typically, the use of screening measures involves not only the collection of but also the review of data to make decisions about which students demonstrate the most risk for poor performance and should be routed for additional intervention. It is unclear whether this investment of resources should be considered when identifying the various ingredients to include when conducting a cost analysis specifically of measurement tools. Future studies might consider whether the use of the measure for screening results in the accurate identification of students who are at risk for failure on other measures of interest. Perhaps the number of students accurately identified by the screener would be a mechanism for determining cost effectiveness ratios for different approaches to screening. This would allow for the examination of cost effectiveness ratios to compare the overall outcomes of the different scores generated by different methods of screening.

Fifth, the automated programs used in the study are free to use (with writeAlizer and ReaderBench also being open source), hence they do not contribute to the total costs. However, in the future, more advanced options might be available for teachers and school districts, such as a secure platform where student performance can be stored and score the writing samples without importing the data in R. Although the automated programs may remain free, these advanced options will likely eventually cost money (e.g., per student, per license). Future cost analyses might consider this possibility in the ingredients (e.g., cost per license, cost-per-student to use the program).

Finally, research regarding the use of automated scores is ongoing and incomplete. As a result, this study examined only a specific range of variables in the sensitivity analyses (e.g., number of students or assessment occasions). As research continues, there will be more variables to examine when investigating the costs associated with the use of this tool as an alternative to traditional WE-CBM. For example, if longer duration samples or an increased number of samples of the same or different genres are required from students, then the administration and hand-scoring time required will increase, as would transcription time required to type the samples for entry into an automated tool. There may be other yet unknown variations related to administration and scoring of the writing samples that



emerge and should be considered in future research. In addition, future studies should examine the utility of having students type their own writing passages as they write. This approach would remove the requirement of transcription after writing samples are generated. Readers might also use the spreadsheet available online to see the extent to which the costs would change as a function of computer use. Interestingly, upon removal of transcription costs, the results of this study might seem rather conservative given that typing would drastically reduce the cost of automated programs but would have little to no impact on the costs associated with hand-calculated WE-CBM scoring. That said, whether generating writing samples via keyboarding and handwriting results in writing samples of equivalent quality is an empirical question. Existing research suggests that variables such as the familiarity a student has with keyboarding impacts the validity of the samples they generate (White et al., 2015). As a result, having students write samples via keyboarding directly should only be used if it can be determined that a sample of equivalent quality and length will be obtained. If keyboarding and handwriting are equivalent, then having students keyboard their samples directly would have a significant positive impact on the cost of implementation given that, based on the current study, transcription costs were the highest when considering the resources required to implement.

## 6. Implications

There are implications for both research and practice as a result of this cost analysis study. As noted above, as features of the measurement approach change, the research on cost must follow suit as these data should be considered as critical as traditional technical adequacy information in evaluating the data generated about a measure. The results of this study present the varying costs associated with the use of traditional WE-CBM implementation and the use of an automated approach. Given the concerning data regarding student writing performance on the national level, tools for effectively and efficiently identifying students with writing difficulties are needed.

Cost analysis studies are critical in the field of school psychology and education more broadly, as cost is an important element of decision-making for implementation. An intervention or innovation may be effective but also very expensive whereas another may be slightly less effective in terms of outcomes but considerably less costly. Thus, the dimension of cost should be an essential element of both research and practice considerations.

More studies that focus on the relative cost of measurement approaches are needed. Selecting and implementing measures for decision-making in schools requires numerous resources, as illustrated here, and should be evaluated for the overall costs associated with their use. Schools engage in many different approaches to measurement or assessment of students, some of which require considerable investments of time. Understanding the opportunity costs associated with the various approaches would provide concrete data to support schools in decision making. The implementation of interventions or measures with the use of university support or through a funded grant will be different from those incurred when a school or district implements independently. Answering questions associated with the costs of measures and interventions under varying conditions is an essential step toward addressing the research to practice gap. This is the true value of examining the cost of an innovation in educational settings.

Similar to when considering interventions, engaging in measurement or assessment activities in schools requires an investment of time and monetary resources and that must be considered in light of the outcomes generated. With intervention research, the outcomes of interest are the changes in student performance that result from the intervention. The examination of cost analysis relative to measures used in school settings should move to consideration of the outcomes that are produced as well. The use of WE-CBM requires a significant investment of time and resources, with most of those resources directed to the process of training personnel for scoring and completing the actual scoring. Examining the resources required to implement an alternative approach to accomplishing these tasks is the first step to understanding the value of using automated programs for scoring. The next step for future studies is to consider the accuracy outcomes of using different measurement approaches for identifying students at risk for poor performance. Additional steps will also involve the creation of a user-friendly interface for teachers to easily generate and use writeAlizer scores in applied settings. Until then, this paper demonstrates the economic advantages of automated scoring approaches over the hand-calculate measures of WE-CBM, as suggested in the best practices outlined by IES (2020).

## 7. Conclusion

The consideration of cost is a relatively new approach to examining the investments of time and other resources in the school setting and examinations of the cost of measurement approaches in the context of universal screening specifically are limited. The results of this study illustrate the economic analysis of various tools for use in screening elementary students in the area of writing including hand-scored WE-CBM as well as the use of automated scoring tools. Cost analysis results indicate that the automated approaches to scoring student writing samples were consistently less expensive when compared to the best performing hand-scored WE-CBM metrics, while offering similar technical adequacy. Ongoing, critical examinations of the costs associated with selecting measures for universal screening as well as the implementation of other educational innovations are needed to inform practical decision-making about adoption and resource allocation.

## Declaration of Competing Interest

We have no known conflict of interest to disclose.

## Acknowledgment

This research was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A190100 awarded to the University of Houston (PI – Milena Keller-Margulis). The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

## References

- Allen, A. A., Jung, P. G., Poch, A. L., Brandes, D., Shin, J., Lembke, E. S., & McMaster, K. L. (2020). Technical adequacy of curriculum-based measures in writing in grades 1–3. *Reading & Writing Quarterly*, 36(6), 563–587. <https://doi.org/10.1080/10573569.2019.1689211>.
- Barrett, C. A., Gadke, D. L., & VanDerHeyden, A. M. (2020). At what cost?: Introduction to the special issue “Return on investment for academic and behavioral assessment and intervention”. *School Psychology Review*, 49(4), 347–358. <https://doi.org/10.1080/2372966X.2020.1817718>.
- Barrett, C. A., Truckenmiller, A. J., & Eckert, T. L. (2020). Performance feedback during writing instruction: A cost-effectiveness analysis. *School Psychology*, 35(3), 193–200. <https://doi.org/10.1037/spq0000356>.
- Barrett, C. A., & VanDerHeyden, A. M. (2020). A cost-effectiveness analysis of classwide math intervention. *Journal of School Psychology*, 80, 54–65. <https://doi.org/10.1016/j.jsp.2020.04.002>.
- 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. <https://doi.org/10.1007/s11145-010-9264-9>.
- Benson, N. F., Floyd, R. G., Kranzler, J. H., Eckert, T. L., Fefer, S. A., & Morgan, G. B. (2019). Test use and assessment practices of school psychologists in the United States: Findings from the 2017 National Survey. *Journal of School Psychology*, 72, 29–48. <https://doi.org/10.1016/j.jsp.2018.12.004>.
- Berninger, V. W., & Winn, W. D. (2006). Implications of advancements in brain research and technology for writing development, writing instruction, and educational evolution. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 96–114). The Guilford Press.
- Crowley, D. M., Dodge, K. A., Barnett, W. S., Corso, P., Duffy, S., Graham, P., Greenberg, M., Haskins, R., Hill, L., Jones, D. E., Karoly, L. A., Kuklinski, M. R., & Plotnick, R. (2018). Standards of evidence for conducting and reporting economic evaluations in prevention science. *Prevention Science*, 19(3), 366–390. <https://doi.org/10.1007/s11211-017-0858-1>.
- Dascaľu, M. (2014). *Analyzing discourse and text complexity for learning and collaborating* (Vol. 534). Springer International Publishing. <https://doi.org/10.1007/978-3-319-03419-5>.
- Deane, P. (2013). Covering the construct: An approach to automated essay scoring motivated by a socio-cognitive framework for defining literacy skills. In M. Shermis, & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 298–312). Routledge.
- Deno, S. L., Mirklin, P. K., & Marston, D. (1980). *Relationships among simple measures of written expression and performance on standardized achievement tests* (Vol. IRLD-RR-22). University of Minnesota, Institute for Research on Learning Disabilities.
- Espin, C., Shin, J., Deno, S. L., Skare, S., Robinson, S., & Benner, B. (2000). Identifying indicators of written expression proficiency for middle school students. *The Journal of Special Education*, 34(3), 140–153. <https://doi.org/10.1177/002246690003400303>.
- Espin, C. A., & Deno, S. L. (2016). Conclusion: Oral reading fluency or reading aloud from text: An analysis through a unified view of construct validity. In K. D. Cummings, & Y. Petscher (Eds.), *The fluency construct: Curriculum-based measurement concepts and applications* (pp. 365–384). Springer. [https://doi.org/10.1007/978-1-4939-2803-3\\_13](https://doi.org/10.1007/978-1-4939-2803-3_13).
- Espin, C. A., Scierka, B. J., Skare, S., & Halverson, N. (1999). Criterion-related validity of curriculum-based measures in writing for secondary school students. *Reading & Writing Quarterly: Overcoming Learning Difficulties*, 15(1), 5–27. <https://doi.org/10.1080/105735699278279>.
- 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. <https://doi.org/10.1007/s11145-015-9550-7>.
- Gansle, K. A., Noell, G. H., Vanderheyden, A. M., Naquin, G. M., & Slider, N. J. (2002). Criterion validity, and time cost of alternate measures for curriculum-based measurement in writing. *School Psychology Review*, 31(4), 477–497. <https://doi.org/10.1080/02796015.2002.12086169>.
- Gansle, K. A., Noell, G. H., Vanderheyden, A. M., Slider, N. J., Hoffpauir, L. D., Whitmarsh, E. L., & Naquin, G. M. (2004). An examination of the criterion validity and sensitivity to brief intervention of alternate curriculum-based measures of writing skill. *Psychology in the Schools*, 41(3), 291–300. <https://doi.org/10.1002/pits.10166>.
- Graham, S., Berninger, V. W., Abbott, R. D., Abbott, S. P., & Whitaker, D. (1997). Role of mechanics in composing of elementary school students: A new methodological approach. *Journal of Educational Psychology*, 89, 170–182. <https://doi.org/10.1037/0022-0663.89.1.170>.
- Hollands, F. M., Hanisch-Cerda, B., Menon, A., Levin, H. M., & Belfield, C. R. (2015). *User manual for CostOut—The CBCSE cost tool kit*. Teachers College: Center for Benefit-Cost Studies of Education, Columbia University. [www.cbccosttoolkit.org](http://www.cbccosttoolkit.org).
- Hollands, F. M., Pratt-Williams, J., & Shand, R. (2021). *Cost analysis standards & guidelines 1.1*. Cost Analysis in Practice (CAP) Project. <https://capproject.org/resources>.
- Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3<sup>rd</sup> ed.). John Wiley & Sons.
- Hosp, M. K., Hosp, J. L., & Howell, K. W. (2016). *The ABCs of CBM: A practical guide to curriculum-based measurement*. The Guilford Press.
- Institute of Education Sciences [IES]. (2020). *Cost analysis: A toolkit* (IES 2020–001). U.S. Department of Education. Retrieved from <https://ies.ed.gov/pubsearch/>.
- Jewell, J., & Malecki, C. K. (2005). The utility of CBM written language indices: An investigation of production-dependent, production-independent, and accurate-production scores. *School Psychology Review*, 34(1), 27–44. <https://doi.org/10.1080/02796015.2005.12086273>.
- Keller-Margulis, M. A., Mercer, S. H., & Matta, M. (2021). Validity of automated text evaluation tools for written-expression curriculum-based measurement: A comparison study. *Reading and Writing: An Interdisciplinary Journal*, 34, 2461–2480. <https://doi.org/10.1007/s11145-021-10153-6>.
- Lai, E. R., Wolfe, E. W., & Vickers, D. (2015). Differentiation of illusory and true halo in writing scores. *Educational and Psychological Measurement*, 75(1), 102–125. <https://doi.org/10.1177/0013164414530990>.
- Leckie, G., & Baird, J.-A. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement*, 48(4), 399–418. <https://doi.org/10.1111/j.1745-3984.2011.00152.x>.
- Levin, H. M., McEwan, P. J., Belfield, C., Bowden, A. B., & Shand, R. (2018). *Economic evaluation in education: Cost-effectiveness and benefit-cost analysis*. SAGE Publications.
- Matta, M., Mercer, S. H., & Keller-Margulis, M. A. (2022). Evaluating validity and bias for hand-calculated and automated written expression curriculum-based measurement scores. *Assessment in Education: Principles, Policy & Practice*. <https://doi.org/10.1080/0969594X.2022.2043240>.
- McMaster, K. L., & Espin, C. (2007). Technical features of curriculum-based measurement in writing: A literature review. *The Journal of Special Education*, 41(2), 68–84. <https://doi.org/10.1177/00224669070410020301>.
- McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.
- Mercer, S. H. (2020). writeAlizer: Generate predicted writing quality and written expression CBM scores (version 1.2.0) [computer software]. <https://github.com/sh Mercer/writeAlizer/>.
- Mercer, S. H., Cannon, J. E., Squires, B., Guo, Y., & Pinco, E. (2021). Accuracy of automated written expression curriculum-based measurement scoring. *Canadian Journal of School Psychology*, 36(4), 304–317. <https://doi.org/10.1177/0829573520987753>.
- 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>.
- National Center for Education Statistics [NCES]. (2012). *The Nation's report card: Writing 2011* (NCES 2012–470). Institute of Education Sciences. Retrieved from <https://nces.ed.gov/nationsreportcard/pdf/main2011/2012470.pdf>.

- 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. <https://doi.org/10.1037/a0013462>.
- Page, E. B. (1966). The imminence of... Grading essays by computer. *The Phi Delta Kappan*, 47(5), 238–243.
- Page, E. B. (2003). Project essay grade: PEG. In M. D. Shermis, & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43–54). Lawrence Erlbaum Associates Publishers.
- Payan, A. M., Keller-Margulis, M., Burrige, A. B., McQuillin, S. D., & Hassett, K. S. (2019). Assessing teacher usability of written expression curriculum-based measurement. *Assessment for Effective Intervention*, 45(1), 51–64. <https://doi.org/10.1177/1534508418781007>.
- Perelman, L. (2014). When “the state of the art” is counting words. *Assessing Writing*, 21, 104–111. <https://doi.org/10.1016/j.asw.2014.05.001>.
- Powell-Smith, K. A., & Shinn, M. R. (2004). *Administration and scoring of written expression curriculum-based measurement (WE-CBM) for use in general outcome measurement*. Edformation, Inc.
- Protopapas, A., & Skaloumbakas, C. (2007). Traditional and computer-based screening and diagnosis of reading disabilities in Greek. *Journal of Learning Disabilities*, 40(1), 15–36. <https://doi.org/10.1177/00222194070400010201>.
- Ritchey, K. D., & Coker, D. L. (2013). An investigation of the validity and utility of two curriculum-based measurement writing tasks. *Reading & Writing Quarterly*, 29(1), 89–119. <https://doi.org/10.1080/10573569.2013.741957>.
- Romig, J. E., Miller, A. A., Therrien, W. J., & Lloyd, J. W. (2020). Meta-analysis of prompt and duration for curriculum-based measurement of written language. *Exceptionality*, 1–17. <https://doi.org/10.1080/09362835.2020.1743706>.
- Romig, J. E., Therrien, W. J., & Lloyd, J. W. (2017). Meta-analysis of criterion validity for curriculum-based measurement in written language. *The Journal of Special Education*, 51(2), 72–82. <https://doi.org/10.1177/0022466916670637>.
- RStudio Team. (2020). *RStudio: Integrated development for R*. PBC: RStudio. <http://www.rstudio.com/>.
- Schneider, M. (2020, April 14). *A new tool to support cost analysis in education research*. Institute of Education Sciences (IES), part of the U.S. Department of Education (ED). Retrieved from <https://ies.ed.gov/director/remarks/4-14-2020.asp>.
- Shapiro, E. S. (2010). *Academic skills problems, fourth edition: Direct assessment and intervention* (4th ed.). The Guilford Press.
- Truckenmiller, A. J., Eckert, T. L., Coddling, R. S., & Petscher, Y. (2014). Evaluating the impact of feedback on elementary aged students’ fluency growth in written expression: A randomized controlled trial. *Journal of School Psychology*, 52(6), 531–548. <https://doi.org/10.1016/j.jsp.2014.09.001>.
- White, S., Kim, Y. Y., Chen, J., & Liu, F. (2015). Performance of fourth-grade students in the 2012 NAEP computer-based writing pilot assessment: Scores, text length, and use of editing tools. In *Working paper series. NCES 2015-119*. National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Retrieved from <https://files.eric.ed.gov/fulltext/ED562627.pdf>.
- Wilson, J. (2018). Universal screening with automated essay scoring: Evaluating classification accuracy in grades 3 and 4. *Journal of School Psychology*, 68, 19–37. <https://doi.org/10.1016/j.jsp.2017.12.005>.
- Wilson, J. (2019). Generalizability of automated scores of writing quality in grades 3–5. *Journal of Educational Psychology*, 111(4), 619–640. <https://doi.org/10.1037/edu0000311>.
- Wilson, J., & Rodrigues, J. (2020). Classification accuracy and efficiency of writing screening using automated essay scoring. *Journal of School Psychology*, 82, 123–140. <https://doi.org/10.1016/j.jsp.2020.08.008>.
- Wilson, J., Roscoe, R., & Ahmed, Y. (2017). Automated formative writing assessment using a levels of language framework. *Assessing Writing*, 34, 16–36. <https://doi.org/10.1016/j.asw.2017.08.002>.
- Wilson, J., & Roscoe, R. D. (2020). Automated writing evaluation and feedback: Multiple metrics of efficacy. *Journal of Educational Computing Research*, 58(1), 87–125.
- Yell, M. L., Deno, S. L., & Marston, D. B. (1992). Barriers to implementing curriculum-based measurement. *Diagnostique*, 18(1), 99–112. <https://doi.org/10.1177/153450849201800109>.