


Evaluating validity and bias for hand-calculated and automated written expression curriculum-based measurement scores

Michael Matta, Sterett H. Mercer & Milena A. Keller-Margulis

To cite this article: Michael Matta, Sterett H. Mercer & Milena A. Keller-Margulis (2022): Evaluating validity and bias for hand-calculated and automated written expression curriculum-based measurement scores, Assessment in Education: Principles, Policy & Practice, DOI: [10.1080/0969594X.2022.2043240](https://doi.org/10.1080/0969594X.2022.2043240)

To link to this article: <https://doi.org/10.1080/0969594X.2022.2043240>

 View supplementary material [↗](#)

 Published online: 28 Feb 2022.

 Submit your article to this journal [↗](#)

 View related articles [↗](#)

 View Crossmark data [↗](#)



Evaluating validity and bias for hand-calculated and automated written expression curriculum-based measurement scores

Michael Matta ^a, Sterett H. Mercer ^b and Milena A. Keller-Margulis ^a

^aDepartment of Psychological, Health & Learning Sciences, University of Houston, Houston, USA;

^bDepartment of Educational and Counselling Psychology, and Special Education, The University of British Columbia, Vancouver, Canada

ABSTRACT

Written expression curriculum-based measurement (WE-CBM) is a formative assessment approach for screening and progress monitoring. To extend evaluation of WE-CBM, we compared hand-calculated and automated scoring approaches in relation to the number of screening samples needed per student for valid scores, the long-term predictive validity and diagnostic accuracy of scores, and predictive and diagnostic bias for underrepresented student groups. Second- to fifth-grade students ($n = 609$) completed five WE-CBM tasks during one academic year and a standardised writing test in fourth and seventh grade. Averaging WE-CBM scores across multiple samples improved validity. Complex hand-calculated metrics and automated tools outperformed simpler metrics for the long-term prediction of writing performance. No evidence of bias was observed between African American and Hispanic students. The study will illustrate the absence of test bias as necessary condition for fair and equitable screening procedures and the importance of future research to include comparisons with majority groups.

ARTICLE HISTORY

Received 29 May 2021



Accepted 9 February 2022


KEYWORDS

Written expression; curriculum-based measurement; automated text evaluation; predictive validity; predictive bias

To effectively adjust writing instruction to meet the needs of struggling writers, educators need assessments that can be used to efficiently screen students and identify those requiring additional supports as well as monitor student writing skill over time. Written expression curriculum-based measurement (WE-CBM) is a formative assessment approach that has been evaluated mostly for use in screening (Ritchey & Coker, 2013) with fewer studies of progress monitoring (e.g. McMaster et al., 2017). In WE-CBM, students write in response to a prompt within a given time, typically three to five min; then, written production is scored for quantitative metrics (McMaster & Espin, 2007).

Research on WE-CBM for screening has focused on several key aspects of reliability and validity, namely identifying specific administration procedures (e.g. task duration and number of writing samples per occasion) and scoring approaches (e.g. hand-calculated vs. automated) that yield optimal estimates of written

CONTACT Michael Matta  mmatta@uh.edu  University of Houston, 403 Farish Hall, 3657 Cullen Blvd., Houston, TX 77204 US

 Supplemental data for this article can be accessed [here](#).

© 2022 Informa UK Limited, trading as Taylor & Francis Group

expression skill. Although research has highlighted the limited reliability of WE-CBM scores estimated from one sample (Keller-Margulis et al., 2016), the number of samples necessary for optimal validity is unclear. Several other key aspects of WE-CBM have not been investigated, such as its ability to predict writing proficiency over long time intervals, the use of automated approaches to generate writing scores, or the potential for predictive bias. The lack of evidence linking WE-CBM scores with long-term outcomes limits the interpretation and use of scores for longitudinal research. Moreover, although automated approaches offer to improve scoring feasibility, there is limited support for the validity and reliability of its scores. Lastly, the use of biased screening assessments might lead to incorrect identification of at-risk students, possibly resulting in adverse consequences for students from underrepresented racial or ethnic groups (Kane, 2013).

The present study addresses three gaps in the literature by evaluating (a) the number of writing samples needed for optimal long-term validity and diagnostic accuracy, (b) the extent to which predictive validity and diagnostic accuracy differ across hand-calculated and automated scoring approaches, and (c) the predictive and diagnostic bias of WE-CBM when used for screening with students of different racial and ethnic groups.

How many screening samples are needed?

Screening with WE-CBM typically involves administration of one writing sample per student per occasion, but this is based on the inference that scores from one sample adequately represent the expected score on all WE-CBM tasks administered under similar circumstances. The underlying assumption is that performance in response to a particular prompt does not differ significantly from alternate prompts. Although alternate-form reliability evidence supports the use of hand-calculated WE-CBM metrics for relative decisions such as rank-ordering students (Marston & Deno, 1981; Weissenburger & Espin, 2005), it would be problematic to rely on such evidence in the context of absolute decisions such as comparing obtained scores to performance standards or monitoring change over time (Keller-Margulis et al., 2016). In fact, moderate to strong correlations between two or more scores obtained for the same metric simply reflect a similar rank order of the students across forms; however, the difference between scores obtained on two or more forms by each student might still be substantial.

To better evaluate alternate-form reliability, scholars have used generalisability theory (Shavelson & Webb, 1991) to determine the optimal number of writing samples per occasion for reliable absolute decisions. Collectively, these studies indicate that reliable scores are unlikely to be obtained using the typical WE-CBM practice of administering a single sample per student per screening occasion (Graham et al., 2016; Keller-Margulis et al., 2016; Kim et al., 2017), with two to seven samples recommended in these studies. The wide range of recommended samples across these studies is in part due to large between-study differences in target student population, writing sample administration time (7–20 min), prompt genre, and scoring metrics that also affect reliability.

Hand-calculated vs. automated WE-CBM scoring

Selecting a specific WE-CBM scoring metric and scoring approach has implications for reliability and validity. Traditionally, WE-CBM writing samples are hand-scored for one or more linguistic metrics, such as counts of total words written (TWW) and words spelled correctly (WSC), the number of sequences of adjacent words spelled correctly and acceptable within the sentence context (correct word sequences, CWS), and the number of correct minus incorrect word sequences (CIWS). In a recent meta-analysis, findings indicated that complex metrics have greater validity than simple metrics (Romig et al., 2017). In particular, CWS and CIWS were better predictors of student performance on criterion writing tests ($r = .51$ and $.60$, respectively) than TWW ($r = .37$) and WSC ($r = .44$). Although they demonstrate greater validity, complex metrics require more time and effort to score than simple metrics, negatively affecting both inter-scorer reliability and scoring feasibility (Amato & Watkins, 2011).

In an effort to improve feasibility, automated approaches have been developed to score screening samples in recent studies with promising results (e.g. Mercer et al., 2019). Contrary to hand-scored approaches, automated approaches employ computer programs to calculate many linguistic indicators for each writing sample and combine these indicators into one or more composite scores of writing quality. Automated scoring considers a broad range of text characteristics, with evidence indicating automated scores adequately represent writing quality on the scored samples.

In a recent study, the writeAlizer R package (Mercer, 2020), a free program that generates overall writing quality scores, was used to score three, 3-min WE-CBM screening samples and its scores were compared to four hand-scored WE-CBM metrics in relation to scores on a standardised writing test in fourth grade (Keller-Margulis et al., 2021). Findings revealed that the average of writeAlizer scores across three screening samples offers validity estimates ($r = .54$ to $.55$) comparable to complex hand-scored metrics ($r = .56$ to $.59$), outperforming simpler hand-scored metrics ($r = .32$ to $.36$) while also improving scoring feasibility. Other studies have compared alternative automated approaches (e.g. Project Essay Grade; Page, 2003) to hand-scored WE-CBM metrics with similar results (Wilson et al., 2016). However, empirical investigations of automated approaches for screening elementary students have only recently emerged and more evidence is needed to support their technical adequacy and use in applied settings.

Predictive validity and diagnostic accuracy of WE-CBM scores

When using WE-CBM scores for screening decisions, educators infer that WE-CBM scores adequately serve as general indicators of writing proficiency. Two types of studies can provide evidence to support this inference. Predictive validity studies allow for the estimation of students' expected scores on a criterion measure, whereas diagnostic accuracy studies are used to predict the degree to which screeners can discriminate between students who reach standards for writing proficiency (Furey et al., 2016).

The development of writing screening measures with good predictive validity and diagnostic accuracy is beneficial for educational practices and applied research. Measures with good short-term predictive validity and accuracy can be used for formative assessment, informing educators about expected student performance on more comprehensive

writing assessments and helping teachers identify struggling writers. Relatedly, writing screening measures with good long-term predictive validity and accuracy can be useful in longitudinal studies investigating cognitive and academic skill development, for example. The long-term predictive validity and diagnostic accuracy of WE-CBM are largely unknown because only one study has investigated them across timeframes longer than one academic year (Espin et al., 1999).

Predictive and diagnostic bias in WE-CBM Screening

Although the predictive validity and diagnostic accuracy of WE-CBM scores has been evaluated, the possibility of predictive and diagnostic bias for underrepresented student groups has not been explored (Evans-Hampton et al., 2002). Predictive and diagnostic bias are defined as systematic errors in the prediction of a criterion measure as a function of group membership that gives an unfair advantage to some groups of students over other groups (American Educational Research Association, National Council on Measurement in Education & American Psychological Association, 2014). Predictive and diagnostic bias that results in disproportionality in either direction (i.e. under-identification or over-identification) would prevent students from receiving appropriate academic supports and possibly contribute to racial disparities in educational outcomes (Hosp et al., 2011; Skiba et al., 2008). Conversely, ensuring screening tools have comparable accuracy across student groups has the potential to reduce racial inequalities and enable students from all backgrounds to receive appropriate academic supports (Betts et al., 2008; Hanushek et al., 2002)

Predictive and diagnostic bias of WE-CBM scores can be evaluated through regression models, in which group membership and scoring metrics are entered as predictors of the criterion measure (Warne et al., 2014). If the effect of group membership is statistically significant, then there is evidence for intercept bias, indicating students with the same score on a WE-CBM metric will be expected to perform systematically better or worse on the outcome measure as a function of their group membership. When the interaction between group and the WE-CBM metric is significant, there is evidence for slope bias, indicating the WE-CBM scores will more accurately predict performance on the criterion measure for one group over others. Both situations affect the degree to which WE-CBM scores accurately predict student performance, leading to overpredictions or underpredictions of student writing proficiency. Additionally, examination of intercept and slope bias can be used to test whether automated WE-CBM scoring approaches might have different accuracy in predicting writing proficiency for students from different backgrounds. The procedures to develop automated scoring models involve the use of machine learning algorithms that are trained to replicate human ratings of writing quality. Therefore, if human ratings show evidence of bias, automated scores might also be biased and amplify pre-existing biases (Amorim et al., 2018).

Ultimately, very little attention has been devoted to exploring possible unintended consequences of using WE-CBM for screening, namely whether validity differs across groups of students that would make screening decisions biased or unfair. Although predictive and diagnostic bias are distinct from fairness and equity, these components influence each other in that biased WE-CBM scores might lead to unfair uses and non-equitable screening decisions. While predictive

and diagnostic bias are determined using statistical procedures, the intended or unintended consequences of WE-CBM score use are evaluated in the light of societal values and social justice (Kline, 2013). In other words, even if WE-CBM is statistically unbiased, it could be deemed unfair based on the consequences of its use in applied settings (Camilli, 2006). In the context of writing screening, examination of the consequences might involve studies of the effects of screening procedures on students identified as needing supplemental or intensive interventions; the allocation of resources within classrooms, schools, and districts; and the design of culturally responsive systems of academic supports (Xu & Drame, 2008). These actions are then considered in connection with any consequences (e.g. placement in special education programmes) that might be harmful to groups of students with certain characteristics.

Purpose of the study

The present study addresses the optimal number of screening samples to administer for hand-scored versus automated text evaluation WE-CBM scoring approaches while examining the differential long-term predictive validity and diagnostic accuracy of WE-CBM across racial and ethnic groups. Specifically, we investigate the extent to which WE-CBM scores collected on five occasions during one academic year for second-through fifth-grade students predict scores on a standardised writing achievement test up to 5 years later, and the extent to which validity and accuracy differ by the number of screening samples and scoring approach used. We address the following primary research questions:

- (1) How does the number of screening samples administered affect the long-term predictive validity and diagnostic accuracy of WE-CBM?
- (2) Are there differences in long-term predictive validity and diagnostic accuracy across WE-CBM scoring approaches, specifically simple hand-scoring, complex hand-scoring, and automated scoring?
- (3) Do WE-CBM scores exhibit predictive or diagnostic bias for students from different racial and ethnic groups?

Method

Participants

The sample included 609 second- to fifth-grade students (299 boys, 49.09% of the sample) from two suburban elementary schools located in the Southwestern United States. Students were similarly distributed across grades. The majority of students were Hispanic (55.66%), followed by African American (33.33%), White (5.91%), and Asian (3.44%). Approximately 8% received special education services. Additional sample demographics are reported in Table 1. Note that the predictive bias analyses were based on African American and Hispanic students only because sample sizes for other groups were too small across every grade ($n < 10$).

Table 1. Sample demographics.

	Second grade	Third grade	Fourth grade	Fifth grade
Sample size	170	144	152	143
Sex boys, <i>n</i> (%)	72 (42)	71 (49)	79 (52)	77 (54)
Race/Ethnicity, <i>n</i> (%)				
African American	60 (35)	43 (30)	54 (36)	46 (32)
Asian	7 (4)	6 (4)	7 (5)	* (≤ 2)
Hispanic	89 (52)	86 (60)	79 (52)	85 (59)
White	12 (7)	9 (6)	8 (5)	7 (5)
Other or Biracial	* (≤ 2)	* (≤ 2)	* (≤ 2)	* (≤ 2)
Special Education, <i>n</i> (%)	9 (5)	10 (7)	12 (8)	17 (12)
English Learners, <i>n</i> (%)	27 (16)	43 (30)	22 (14)	* (≤ 2)
Eligible for free or reduced-price meals, <i>n</i> (%)	119 (70)	103 (72)	100 (73)	98 (68)
Gifted, <i>n</i> (%)	13 (8)	23 (16)	15 (10)	14 (10)

Note. Per the Texas Administrative Code, the number of students from small groups was masked to maintain confidentiality.

Measures

Hand-scored WE-CBM

Writing samples were hand-scored for four WE-CBM metrics capturing simple or more complex aspects of written expression. TWW and WSC are simple metrics because they do not consider the syntactic and semantic context of the sentence. TWW is the total number of words written with ‘word’ defined as any letter or group of letters delimited by white spaces. WSC is the count of words correctly spelled without considering context. By contrast, CWS and CIWS are more complex metrics because scoring considers the within-sentence and between-sentence context. CWS is the number of correct sequences of two adjacent words separated by a space or punctuation; this metric considers whether words are spelled correctly and placed within proper syntactic and semantic contexts. Finally, CIWS is the difference between CWS and the number of incorrect sequences.

Interrater reliability was calculated on a subset of randomly selected writing samples (41.66% of the total) balanced across grades and time points. Concordance correlation coefficients indicated interrater reliability was .99, 95% CI [.98, .99] for TWW; .98, 95% CI [.98, .99] for WSC; .96, 95% CI [.95, .96] for CWS; and .87, 95% CI [.85, .88] for CIWS. Agreement was within the almost perfect range ($\rho = .81$ to 1) for all metrics based on qualitative descriptors that were developed for similar measures of observer agreement (Landis & Koch, 1977).

Automated WE-CBM

We used the writeAlizer R package (Mercer, 2020) to generate writing quality scores from the output of two different text complexity analysis tools, Coh-Metrix (McNamara et al., 2014) and ReaderBench (Dascălu, 2014). Coh-Metrix and ReaderBench, both of which are freely accessible online, use natural language processing techniques to generate many indices reflecting lexical, syntactic, semantic, and discourse features of written language originally designed to predict text readability. The writeAlizer scoring model is an ensemble of seven machine learning algorithms that were trained on 7-min narrative writing samples from second- to fifth-grade students (see, Mercer et al., 2019). The weightings of Coh-Metrix and ReaderBench indices in each algorithm and in the overall models are available on the writeAlizer

GitHub site. Prior research indicates Coh-Metrix indices, when combined in latent factors representing word-level, sentence-level, and discourse-level writing skills (Wilson et al., 2017), can predict performance on a standardised writing achievement test for sixth- and eighth-grade students. Similarly, our prior work has shown writeAlizer-generated writing quality scores based on Coh-Metrix and Readerbench indices can predict performance on a standardised writing assessment for fourth-grade students (Keller-Margulis et al., 2021).

STAAR writing test

The State of Texas Assessments of Academic Readiness (STAAR) Writing Test is a criterion-referenced test used for the evaluation of writing proficiency at the end of the academic year for students in fourth and seventh grade. The test was administered annually from the 2011–2012 to the 2020–2021 school year, with the exception of 2019–2020 due to the Covid-19 pandemic. The Texas Education Agency (TEA), the state agency that oversees primary and secondary public education, makes available all administered test forms on its website after administration. Initially, the test was administered in two 4-h sessions, however, beginning in 2015–2016 it was shortened to one, 4-h session with a reduced number of items. Students write one or two essays and make corrections and revisions to a number of stories by answering multiple-choice questions. The essay was scored on a scale from 1 to 4 by two independent, trained raters who evaluated the essay on performance criteria for three dimensions (i.e. text organisation quality, idea development, and use of linguistic conventions) on a holistic rubric. There were 24 multiple-choice questions, with each correct answer giving one point. Classroom teachers administer the test following a standardised script, and students can complete tasks in the order they prefer.

TEA is responsible for the development and validation of the STAAR writing test and sets performance standards through a nine-step process involving a series of linking studies. The process is designed to provide empirical support for the construct validity of STAAR scores by estimating the relationship of student performance with existing writing measures; for example, there was a strong correlation between student writing performance on the STAAR test in fourth and seventh grade ($r = .62$) as well as with the scores obtained on the ReadiStep ($r = .63$) and EXPLORE ($r = .66$), two tests of academic achievement that are linked to the Scholastic Aptitude Test (SAT) and the American College Testing (ACT; Texas Education Agency, 2013).

Study design and procedures

Teachers collected WE-CBM screening samples in their classrooms. Students completed one, 3-min narrative WE-CBM task in response to age-appropriate story starters at five time-points equally spread across the 2011–2012 school year. Story starters were different across grades and time points. At the end of each session, teachers completed a checklist to ensure administration fidelity. All administrations returned a perfect implementation. Subsequent to initial WE-CBM data collection in 2011–2012, students also completed the STAAR writing test in fourth and seventh grade. Specifically, students in fourth grade

during the 2011–2012 school year completed the STAAR test within the same year and three years later, third-grade students after one and four years, and so on. Fifth-grade students were the only group to complete the STAAR test in seventh grade only.

A university Institutional Review Board (IRB) approved use of the deidentified WE-CBM data for research purposes. Trained graduate students hand-scored writing samples for four WE-CBM metrics and transcribed the text into an electronic format. A postdoctoral fellow proficient in R processed all the transcribed text files through Coh-Metrix and ReaderBench, with their output files processed by the writeAlizer R package to generate writing quality scores.

STAAR writing data were obtained with approval of and in collaboration with the Education Research Center (ERC), a state-authorized data warehouse that holds individual-level data of students enrolled in primary and secondary schools in Texas and can match data with those collected by TEA and other state agencies through secured systems while maintaining confidentiality. Our team worked with the ERC and TEA to link students who completed the five WE-CBM tasks in 2012 with STAAR writing test results in fourth and seventh grade administered between 2011–2012 and 2016–2017.

Data analysis

Analyses were conducted in RStudio (RStudio Team, 2020). Of the 722 students who originally completed WE-CBM tasks, 113 could not be linked to STAAR data due to unavailable linking identification numbers in the TEA archive and were therefore excluded. For the 609 matched students, some data were missing due to student absences on one or more of the 5 days of WE-CBM data collection. WE-CBM tasks were not administered in mid-spring to students enrolled in one of the two campuses, hence the rate of missing data for that time point was substantially higher (45.39% to 52.94%) than other administrations (2.80% to 25.29%). Missingness of STAAR writing scores ranged from 2.10% to 15.29% with no clear pattern of attrition over time. Assuming these data were missing at random, we generated 1000 multiply imputed datasets in Mplus (Muthén & Muthén, 1998–2017) and imported them in RStudio via the miceadds package (Robitzsch & Grund, 2021). This procedure, known as multiple imputation by chained equations, makes use of all the variables in the data set to predict missing values and returns a complete dataset. Instead of conducting the analyses on one data set only, a large number of imputed data sets is typically generated each of which contain slightly different imputed values that ultimately allows to reduce the bias associated with the missing data. Analyses were conducted on each imputed dataset separately then pooled together using Rubin's rules via the mice (Van Buuren & Groothuis-Oudshoorn, 2011) and psfmi (Heymans & Eekhout, 2021) packages. Rubin's rule combines the results accounting both for the variance within imputed each data set (i.e. the parameter of uncertainty in inferential statistical models) and across data sets (i.e. the variation across the same parameter, such as correlation coefficients, estimated for each data set).

To assess the number of screening samples needed to produce optimal predictive validity and diagnostic accuracy (Research Question #1), we conducted analyses on three sets of WE-CBM scores in relation to STAAR writing test performance: a) spring WE-CBM scores, b) averaged scores across fall, winter, and spring time points (typical administration times for universal screening), and c) averaged scores across five time

Table 2. Predictive validity for written expression CBM scores with the standardised writing scores.

Measure	<i>n</i>	Same year			One year			Two years ¹			Two years ²			Three years			Four years			Five years		
		<i>r</i>	LL	UL	<i>r</i>	LL	UL	<i>r</i>	LL	UL	<i>r</i>	LL	UL	<i>r</i>	LL	UL	<i>r</i>	LL	UL	<i>r</i>	LL	UL
TWW	1	.37	.21	.51	.29	.13	.44	.27	.11	.42	.11 ^{ns}	-.06	.27	.25	.09	.40	.23	.06	.38	.28	.11	.44
	3	.37	.22	.50	.39	.23	.53	.39	.24	.51	.22	.06	.37	.22	.06	.37	.31	.14	.46	.37	.22	.50
	5	.38	.23	.51	.37	.21	.52	.43	.28	.55	.30	.14	.45	.27	.12	.42	.31	.14	.47	.37	.22	.50
WSC	1	.40	.24	.53	.35	.19	.49	.30	.14	.45	.16	-.01	.32	.28	.12	.43	.28	.11	.43	.32	.15	.47
	3	.41	.26	.54	.44	.29	.58	.44	.29	.56	.30	.15	.45	.27	.11	.41	.35	.19	.50	.43	.29	.55
	5	.41	.26	.54	.44	.28	.57	.48	.35	.60	.39	.24	.52	.31	.15	.45	.35	.19	.50	.42	.28	.55
CWS	1	.48	.34	.60	.39	.24	.53	.38	.22	.51	.29	.13	.44	.41	.27	.54	.35	.19	.49	.41	.26	.55
	3	.58	.46	.68	.51	.36	.63	.49	.36	.61	.42	.27	.55	.45	.31	.57	.41	.26	.55	.49	.35	.60
	5	.57	.44	.68	.50	.36	.62	.55	.43	.66	.51	.37	.62	.47	.33	.59	.41	.26	.55	.49	.35	.60
CIWS	1	.46	.32	.58	.38	.23	.52	.40	.24	.53	.39	.23	.52	.47	.33	.59	.37	.22	.51	.44	.29	.57
	3	.60	.48	.70	.52	.38	.63	.49	.35	.60	.48	.35	.60	.51	.38	.62	.43	.28	.56	.49	.35	.60
	5	.60	.48	.70	.53	.39	.64	.57	.45	.67	.54	.41	.65	.53	.39	.64	.44	.30	.57	.51	.37	.62
wA:CM	1	.47	.33	.60	.44	.29	.57	.39	.23	.52	.31	.15	.46	.42	.27	.55	.40	.25	.53	.43	.28	.57
	3	.61	.49	.70	.55	.41	.66	.52	.39	.63	.47	.32	.59	.46	.32	.58	.44	.29	.57	.51	.37	.62
	5	.64	.53	.73	.53	.38	.65	.58	.45	.68	.53	.39	.64	.51	.38	.63	.47	.32	.60	.54	.41	.64
wA:RB	1	.47	.33	.60	.48	.34	.60	.34	.19	.48	.30	.14	.44	.35	.20	.49	.42	.27	.55	.41	.26	.55
	3	.56	.44	.67	.53	.39	.65	.50	.37	.61	.45	.31	.57	.41	.27	.54	.45	.29	.58	.50	.37	.61
	5	.56	.44	.67	.51	.36	.64	.56	.44	.67	.56	.43	.66	.46	.32	.58	.44	.28	.58	.52	.39	.63

Note. *n* = number of samples; LL = Lower Limit; UL = Upper Limit; TWW = Total Words Written; WSC = Words Spelled Correctly; CWS = Correct Word Sequences; CIWS = Correct Minus Incorrect Word Sequences; wA:CM = writeAlizer based on Coh-Matrix scores; wA:RB = writeAlizer based on ReaderBench scores.

¹Second-grade students completing the STAAR test in fourth grade. ²Fifth-grade students completing the STAAR test in seventh grade.

All correlations are statistically significant at $p < .05$ except where indicated by *ns*.

points (adding mid-Fall and mid-Spring). Descriptive statistics of WE-CBM and STAAR writing scores by group are located in the online supplemental materials (see Table S1 and S2). Predictive validity was evaluated using Pearson's *r*, and for diagnostic accuracy, we used WE-CBM scores to predict a binary variable indicating proficiency on the STAAR writing test. Although the proficiency threshold varied across years, students who did not reach the standard generally performed below the 30th percentile. The Area Under the Curve (AUC) served as an overall diagnostic accuracy statistic and was calculated via the pROC package (Robin et al., 2011). We used a modified version of the Rubin's rule to pool the AUC coefficients across the 1000 imputed datasets (Licht, 2010).

Once the optimal number of screening samples was determined based on the strongest coefficients with the state test, validity and diagnostic accuracy were evaluated against absolute criteria and relative to one another (Research Question #2). In terms of absolute criteria, we considered acceptable validity for Pearson's *r* above .50, a commonly used threshold in the context of WE-CBM (McMaster & Campbell, 2008); we also considered AUC values above .95 as indicators of excellent accuracy, values between .85–.95 as very good, values between .75–.85 as reasonable, and below .75 as inadequate (Smolkowski et al., 2016). In terms of relative comparisons, we conducted a series of multiple contrasts to estimate whether there were statistically significant differences between pairs of scoring metrics. For one group of students, we also were able to compare validity and accuracy coefficients of each metric with the ability of scores on the fourth-grade STAAR test to predict students' STAAR scores three years later in seventh grade. STAAR coefficients may offer a reasonable upper bound for interpreting WE-CBM coefficients given that prior performance on the same test is likely to be the best predictor of

Table 3. Intercept and slope bias for predicting standardised writing test scores across Hispanic and African American students.

Metrics	term	Same year			One year			Two years ¹			Two years ²			Three years			Four years			Five years		
		H	AA	p	H	AA	p	H	AA	p	H	AA	p	H	AA	p	H	AA	p	H	AA	p
		b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b
TWW	Intercept	21.91	16.59	.15	22.89	21.19	.63	20.11	19.42	.81	28.74	31.41	.70	33.84	35.73	.79	27.19	25.58	.69	22.99	24.23	.68
	Slope	0.14	0.20	.55	0.16	0.16	.98	0.25	0.26	.92	0.31	0.28	.83	0.25	0.16	.61	0.14	0.16	.89	0.25	0.18	.57
WSC	Intercept	21.74	16.16	.11	22.27	19.88	.45	20.32	18.84	.55	25.74	28.12	.71	33.35	33.86	.94	26.94	24.67	.53	23.02	23.44	.87
	Slope	0.16	0.23	.47	0.19	0.22	.77	0.28	0.34	.59	0.40	0.37	.87	0.28	0.22	.76	0.16	0.21	.69	0.28	0.25	.81
CWS	Intercept	19.80	15.20	.08	22.22	19.88	.34	20.78	18.39	.24	25.33	25.87	.91	30.09	30.26	.97	26.67	24.82	.52	23.30	22.91	.86
	Slope	0.25	0.31	.50	0.22	0.26	.68	0.32	0.49	.14	0.45	0.48	.83	0.43	0.39	.79	0.20	0.24	.69	0.33	0.38	.70
CIWS	Intercept	22.88	19.95	.04	24.54	22.65	.18	24.32	23.42	.43	32.96	32.14	.78	35.46	35.42	.99	28.76	27.49	.45	26.92	26.69	.85
	Slope	0.25	0.27	.79	0.22	0.23	.88	0.30	0.47	.10	0.35	0.43	.40	0.43	0.39	.79	0.19	0.21	.88	0.32	0.39	.52
wA:CM	Intercept	32.65	30.96	.30	35.20	36.27	.73	36.81	41.25	.26	49.31	53.53	.15	51.38	51.42	.99	38.85	41.46	.46	40.12	41.55	.72
	Slope	0.02	0.03	.45	0.02	0.03	.37	0.02	0.03	.24	0.03	0.05	.32	0.03	0.04	.73	0.02	0.03	.32	0.02	0.02	.75
wA:RB	Intercept	33.07	31.64	.51	34.74	37.34	.43	38.41	44.11	.22	51.44	53.88	.41	52.79	51.97	.84	38.09	41.70	.34	42.07	43.35	.79
	Slope	0.02	0.02	.47	0.02	0.02	.21	0.02	0.03	.23	0.04	0.04	.67	0.03	0.03	.95	0.01	0.02	.24	0.02	0.02	.87

Note. H = Hispanic students; AA = African American students; LL = Lower Limit; UL = Upper Limit; TWW = Total Words Written; WSC = Words Spelled Correctly; CWS = Correct Word Sequences; CIWS = Correct Minus Incorrect Word Sequences; wA:CM = writeAlizer based on Coh-Metrix scores; wA:RB = writeAlizer based on ReaderBench scores.

¹Second-grade students completing the STAAR test in fourth grade. ²Fifth-grade students completing the STAAR test in seventh grade.

subsequent performance. We used the *cocor* R package (Diedenhofen et al., 2015) to test differences in predictive validity and the *pROC* R package for differences in diagnostic accuracy.

Finally, to conduct the predictive and diagnostic bias analyses (Research Question #3), we first estimated multiple logistic regression models to test the differences in the regression intercepts and slopes between Hispanic and African American students via the *stats* package. We entered the dummy-coded term of group membership and its interaction with WE-CBM metrics in the model to predict the scores on the STAAR writing test. Then, we calculated separate AUC values for Hispanic and African American students and evaluated whether differences between the two groups were significant based on logistic regressions.

Results

Research question #1: number of samples for optimal validity and accuracy

Validity coefficients (r) and AUC values of hand-calculated and automated scores for prediction of STAAR writing scores along with their 95% confidence intervals are reported in Tables 2 and 3. For each metric, coefficients were calculated using scores from one writing sample (i.e. Spring 2012) and the average of three and five samples.

Overall, three key findings emerged. First, validity and AUC coefficients increased as a function of the number of writing samples. For instance, *writeAlizer:Coh-Metrix* (CM) had r coefficients of .47 (95% CI [.33, .60]) and AUC of .80 (95% CI [.69, .87]) for one sample, r of .61 (95% CI [.49, .70]) and AUC of .82 (95% CI [.74, .89]) for the average of three samples, and r of .64 (95% CI [.53, .73]) and AUC of .85 (95% CI [.76, .90]) for five samples in relation to STAAR test within the same year. This pattern was consistent across metrics and years between the administration of WE-CBM and the STAAR writing test.

Second, validity and accuracy coefficients increased from one to three samples across metrics and years. For instance, the within-year coefficients for CIWS increased from r of .46 (95% CI [.32, .58]) to .60 (95% CI [.48, .70]) and AUC increased from .80 (95% CI [.71, .86]) to .86 (95% CI [.79, .92]) when based on one vs. three samples. Validity and accuracy coefficients were equal or showed modest gains from three to five writing samples. For instance, CIWS did not show improvements from three to five samples within the same year, but increased from .49 (95% CI [.35, .60]) to .57 (95% CI [.45, .67]) in relation to the writing test completed two years later. Given these results, we used the average across five time points in subsequent analyses to maximise the validity of scores for prediction of STAAR writing test performance.

Third, the magnitude of the coefficients slowly attenuated over time, however, this might not be a simple function of time given the differences were small, and change did not follow a clear linear trend. For example, validity coefficients for CWS were .57 (95% CI [.44, .68]) within the same year and .49 (95% CI [.35, .60]) five years later with the latter being negligibly greater than the coefficient for the previous two years ($r = .47$, 95% CI [.33, .59] and $r = .41$, 95% CI [.26, .55]).

Research question #2: hand-Calculated vs automated scores

Predictive validity

Table 2 displays the validity coefficients for hand-calculated and automated scoring approaches.¹ As an expected upper bound when interpreting the magnitude of the WE-CBM validity coefficients, the student performance on the STAAR test in fourth grade was highly correlated ($r = .67$, 95% CI [.57, .75]) with STAAR test scores three years later. Overall, two key findings can be observed.

First, complex hand-calculated and automated scores had $r \geq .50$. CWS and writeAlizer:RB had coefficients above the cut-off up to 2 years with writeAlizer:RB exceeding the cut-off also at five years. CIWS and writeAlizer:CM had coefficients greater than .50 across every time point except at four years.

Second, more complex hand-scored WE-CBM metrics showed higher validity coefficients than simple WE-CBM scores. No differences were observed among the validity coefficients of CWS, CIWS, writeAlizer:CM, and writeAlizer:RB for prediction of the STAAR writing test over time (see Table S3). Among the simple hand-scored WE-CBM metrics, WSC had substantially higher validity coefficients than TWW.

Diagnostic accuracy

Table 3 illustrates the diagnostic accuracy of hand-calculated and automated scores over time. For reference when interpreting the diagnostic accuracy of WE-CBM scores, the diagnostic accuracy of STAAR in fourth grade for identifying whether students would pass the STAAR test three years later was reasonable to very good (AUC = .85, 95% CI [.77, .90]). Overall, the patterns for diagnostic accuracy were similar to the validity coefficients. First, complex hand-calculated and automated scores within the same year showed AUC values above .75, a threshold for identifying reasonable academic screeners. Additionally, CIWS and writeAlizer:CM had AUC values greater than .85 indicating very good diagnostic accuracy. Generally, complex hand-calculated and automated scores displayed reasonable accuracy for several years after administration of WE-CBM. For example, CIWS had AUC values between .75 and .85 up to three years later. By contrast, simple metrics (i.e. TWW and WSC) had AUC values below .75, indicating poor diagnostic accuracy.

Second, simple WE-CBM metrics showed significantly lower AUC values than complex hand-calculated and automated metrics. Although the size of the differences varied, the pattern was consistent over time (see Table S4). For instance, writeAlizer:CM had an AUC value of .85 (95% CI [.76, .90]) within the year and .82 (95% CI [.73, .89]) one year later, whereas TWW had an AUC value of .71 (95% CI [.60, .80]) and .76 (95% CI [.65, .84]), respectively. There were no differences among complex hand-calculated metrics and automated scores.

Research question #3: predictive and diagnostic bias

The results displayed in Table 4 shows no indication of intercept or slope bias between Hispanic and African American students. Table 5 also indicates that AUC values were not statistically different between the groups.

Table 4. Diagnostic accuracy for predicting reaching proficiency standards on the standardised writing test.

Metrics	<i>n</i>	Same year			One year			Two years ¹			Two years ²			Three years			Four years			Five years		
		AUC	LL	UL	AUC	LL	UL	AUC	LL	UL	AUC	LL	UL	AUC	LL	UL	AUC	LL	UL	AUC	LL	UL
TWW	1	.74	.63	.83	.68	.55	.78	.66	.55	.76	.50	.39	.61	.69	.58	.78	.72	.58	.82	.60	.47	.71
	3	.71	.60	.80	.78	.67	.85	.71	.60	.79	.56	.44	.67	.64	.53	.74	.73	.61	.83	.65	.54	.74
	5	.71	.60	.80	.76	.65	.84	.73	.63	.81	.59	.47	.69	.65	.55	.75	.73	.61	.82	.65	.54	.75
WSC	1	.76	.65	.85	.70	.57	.81	.67	.55	.76	.53	.42	.65	.71	.60	.80	.73	.59	.84	.61	.48	.73
	3	.74	.63	.82	.79	.69	.87	.72	.62	.81	.61	.50	.71	.67	.56	.76	.73	.61	.83	.67	.56	.76
	5	.73	.63	.82	.78	.68	.86	.75	.65	.83	.65	.54	.75	.68	.57	.77	.74	.61	.83	.68	.57	.77
CWS	1	.80	.71	.87	.70	.56	.81	.67	.56	.76	.62	.51	.72	.78	.68	.86	.73	.59	.84	.65	.52	.75
	3	.83	.74	.89	.79	.69	.87	.75	.65	.83	.68	.56	.78	.77	.66	.84	.75	.63	.84	.70	.59	.79
	5	.83	.74	.89	.79	.69	.87	.79	.69	.86	.72	.61	.81	.76	.66	.84	.75	.63	.84	.71	.60	.80
CIWS	1	.80	.71	.86	.66	.53	.77	.64	.53	.74	.66	.55	.76	.80	.69	.87	.70	.57	.81	.68	.55	.78
	3	.86	.79	.92	.75	.64	.84	.73	.63	.81	.71	.59	.81	.80	.70	.87	.72	.60	.82	.71	.60	.80
	5	.87	.79	.92	.77	.67	.85	.78	.69	.86	.75	.63	.83	.79	.69	.86	.74	.62	.83	.73	.62	.82
wA:CM	1	.80	.69	.87	.73	.62	.82	.69	.58	.78	.64	.52	.74	.75	.65	.83	.69	.56	.80	.69	.57	.79
	3	.82	.74	.89	.82	.73	.89	.76	.66	.83	.68	.56	.77	.76	.66	.84	.70	.58	.80	.71	.61	.80
	5	.85	.76	.90	.82	.73	.89	.78	.69	.86	.71	.60	.80	.78	.68	.85	.74	.62	.83	.74	.63	.82
wA:RB	1	.79	.68	.86	.78	.66	.86	.68	.58	.77	.63	.52	.73	.75	.64	.83	.76	.63	.86	.66	.53	.77
	3	.81	.71	.87	.83	.74	.90	.76	.66	.84	.68	.56	.77	.76	.65	.84	.76	.64	.85	.70	.60	.79
	5	.82	.73	.88	.83	.73	.89	.78	.69	.86	.73	.62	.81	.77	.67	.84	.76	.64	.85	.72	.61	.80

Note. *n* = number of samples; AUC = Area Under the Curve, LL = Lower Limit; UL = Upper Limit; 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.

¹Second-grade students completing the STAAR test in fourth grade. ²Fifth-grade students completing the STAAR test in seventh grade.

Discussion

This study examined several key issues associated with WE-CBM as a formative assessment of writing performance and expands the literature in three directions. First, WE-CBM scores calculated from one writing sample were associated with validity coefficients consistently below .50 and accuracy below .80. Validity coefficients of TWW, WSC, and CWS within the same year align with results of a recent meta-analysis on hand-calculated WE-CBM metrics (Romig et al., 2017). By contrast, we found lower validity coefficients for CIWS ($r = .46$ vs $r = .65$ in Romig et al., 2017). Averaging scores across three and five writing samples improved coefficients, especially from one to three samples, with marginal improvements from three to five samples. Predictive validity coefficients indicated that WE-CBM scores are stable predictors of students' expected writing performance over time. However, diagnostic accuracy results suggest that the use of WE-CBM scores might be limited within the same year for screening decisions. CIWS and writeAlizer in combination with Coh-Metrix showed coefficients with very good accuracy, while CWS and writeAlizer in combination with ReaderBench coefficients with reasonable accuracy. Given that no measure showed very good accuracy beyond the same year, we do not recommend the use of WE-CBM scores to make predictions for the student performance on the STAAR test.

Second, we found complex hand-scored and automated WE-CBM scores showed improved predictive validity and diagnostic accuracy compared to simple hand-scored WE-CBM metrics. Using the average across five samples, complex hand-calculated and automated scores showed reasonable to very good diagnostic accuracy. Conversely,



Table 5. Tests of diagnostic bias across Hispanic and African American students.

Metrics	Same year			One year			Two years ¹			Two years ²			Three years			Four years			Five years		
	H	AA	p	H	AA	p	H	AA	p	H	AA	p	H	AA	p	H	AA	p	H	AA	p
TWW	.71	.67	.72	.83	.62	.52	.80	.63	.46	.60	.89	.69	.51	.17	.77	.77	.66	.91	.73	.53	.66
WSC	.71	.71	.97	.84	.67	.65	.80	.68	.76	.68	.85	.70	.56	.25	.76	.69	.69	.95	.73	.54	.79
CWS	.78	.83	.52	.84	.71	.87	.79	.78	.43	.74	.73	.76	.70	.51	.78	.72	.72	.87	.73	.65	.66
CIWS	.77	.92	.10	.80	.74	.89	.72	.80	.13	.75	.29	.77	.80	.83	.76	.72	.72	.95	.69	.73	.43
wA:CM	.87	.81	.77	.89	.76	.87	.83	.74	.85	.74	.86	.77	.75	.91	.78	.70	.70	.84	.75	.71	.88
wA:RB	.81	.78	.99	.87	.77	.70	.81	.76	.73	.74	.85	.77	.71	.52	.79	.74	.74	.66	.75	.64	.80

Note: AUC = Area Under the Curve; TWW = Total Words Written; WSC = Words Spelled Correctly; CWS = Correct Word Sequences; CIWS = Correct Minus Incorrect Word Sequences; writeAlizer:CM = writeAlizer based on Coh-Matrix scores; writeAlizer:RB = writeAlizer based on ReaderBench scores.

¹Second-grade students completing the STAAR test in fourth grade. ²Fifth-grade students completing the STAAR test in seventh grade. Reported p-values are for the interaction terms of the logistic regression models.

regardless of the number of samples, simple hand-calculated metrics displayed insufficient validity and poor diagnostic accuracy. This pattern was stronger for prediction of writing outcomes within the same year.

Third, neither predictive nor diagnostic bias were observed for WE-CBM scores between African American and Hispanic students. Given that the absence of test bias is a necessary condition for fair and equitable screening procedures, the use of hand-calculated and automated WE-CBM scores might not lead to differential predictions across these two racial and ethnic groups or such differences might be of a small magnitude. However, lack of statistical evidence for predictive or diagnostic bias does not guarantee fair interpretation or use of WE-CBM scores in screening (Kline, 2013). Predictive bias is narrowly defined as a psychometric property of test scores, whereas fairness and equity refer to use of scores in practice to make decisions with consequences for students in terms of opportunities and support (Kunnan, 2000). For more than 50 years, researchers have consistently shown that students from underrepresented racial or ethnic groups in the U.S. obtain lower test scores (Persky et al., 2003) and are disproportionately represented in special education programmes (Skiba et al., 2016). If the use of a statistically unbiased screener still results in differential access to academic supports for students as a function of group membership (e.g. racial or ethnic group), then the use of WE-CBM for screening might still be unfair. When WE-CBM scores are used for decision-making, educators draw additional inferences that might affect student support and learning. These inferences linking the use of test scores (e.g. to determine unsatisfactory performance) to educational decisions will need further investigation to ensure the implementation of fair practices for students from all racial and ethnic groups (Kane, 2013).

Limitations and future directions

Our findings should be interpreted in the light of three limitations. First, different forms of the STAAR writing test were used over the five years considered. However, several factors minimise possible effects of changing test forms on our findings: (a) each form had the same three-subtest structure (i.e. written composition, editing, and revising), (b) each form had evidence of similar external validity through strong correlations with STAAR reading scores ($r = .73$ to $.78$), and (c) each form completed in fourth grade showed strong temporal stability with the form administered to students in seventh grade ($r = .68$ to $.72$).

Second, predictive and diagnostic bias results are limited by the use of race and ethnicity as grouping variables. Although this information was provided by the state education agency, hence deemed reliable, it is important to stress that race and ethnicity are not discrete categories but rather complex social constructs varying on a spectrum and influenced by social and cultural identities (Han et al., 2019). For instance, students from the same racial group might develop different identities as a function of how multiple variables (e.g. role of religion in the family or cultural diversity in friendship) interact with their cultural heritage. Future studies should explore race and ethnicity as dimensional constructs instead of discrete categories and consider the intersectionality of race and ethnicity with other relevant demographic characteristics (e.g. gender, socioeconomic status, and language proficiency).

Third, the investigation of predictive and diagnostic bias did not include White students because their group size was not sufficient at any time point. Given the absence of the majority group in the analyses, the investigation of test bias was not optimal to determine whether WE-CBM scores over- or under-estimated the performance of students from historically marginalised groups. Future studies should sample students across a larger geographic area to better enable tests of predictive and diagnostic bias across multiple groups and improve generalisation of the results. The lack of evidence for test bias between African American and Hispanic students still leaves open the possibility that WE-CBM scores might be biased against both when compared with other racial and ethnic groups (Keegan et al., 2013).

Note

1. Separate multilevel regression models were estimated with each metric as the outcome for the first two time points to examine the effects of a nested structure (i.e. level-2 of classroom) on the findings of the study. The Intra-Class Coefficients (ICC) were substantial for the random intercepts across models ($ICC > .10$) but negligible for the random slopes ($ICC < .10$). In other words, while high variability was observed across the intercepts of classrooms, the same variability was not observed for the slopes. Thus, the results presented in this section and the inferences drawn in the Discussion would not change after the inclusion of the effects of classrooms.

Acknowledgments

This research was supported by the Society for the Study of School Psychology (SSSP) Early Career Research Award. This research was also 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.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributors

Michael Matta, Ph.D., is a Research Scientist in School Psychology at the University of Houston. His research focuses on the validation of brief, computer-based assessments to extend the interpretation of student test scores to real-world outcomes and help develop fair and equitable ways to use such scores in applied settings.

Sterett H. Mercer, Ph.D., is a Professor in Special Education in the Department of Educational and Counselling Psychology & Special Education, at the University of British Columbia. His research focuses on the measurement of student academic skills in response to instruction and intervention. <https://ecps.educ.ubc.ca/person/sterett-mercero/>

Milena A. Keller-Margulis, Ph.D., is an Associate Professor of School Psychology in the Psychological, Health, and Learning Sciences Department at the University of Houston. Her research focuses on the use of curriculum-based measures to assess academic skills in the context of multi-tiered systems of support. <https://www.uh.edu/education/about/directory/employee-profile/index.php?id=504>

ORCID

Michael Matta  <http://orcid.org/0000-0003-4266-0130>

Sterett H. Mercer  <http://orcid.org/0000-0002-7940-4221>

Milena A. Keller-Margulis  <http://orcid.org/0000-0001-7539-5375>

References

- Amato, J. M., & Watkins, M. W. (2011). The predictive validity of CBM writing indices for eighth-grade students. *The Journal of Special Education, 44*(4), 195–204. <https://doi.org/10.1177/0022466909333516>
- American Educational Research Association, National Council on Measurement in Education, & American Psychological Association. (2014). *Standards for educational and psychological testing*.
- Amorim, E., Cançado, M., & Veloso, A. (2018). Automated essay scoring in the presence of biased ratings. In *Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers)* (New Orleans, USA: Association for Computational Linguistics), 229–237. <https://doi.org/10.18653/v1/N18-1021>
- Betts, J., Reschly, A., Pickart, M., Heistad, D., Sheran, C., & Marston, D. (2008). An examination of predictive bias for second grade reading outcomes from measures of early literacy skills in kindergarten with respect to english-language learners and ethnic subgroups. *School Psychology Quarterly, 23*(4), 553–570. <https://doi.org/10.1037/1045-3830.23.4.553>
- Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational Measurement* (pp. 221–256). Praeger.
- Dascălu, 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>
- Diedenhofen, B., Musch, J., & Olivier, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PLoS ONE, 10*(4), e0121945. <https://doi.org/10.1371/journal.pone.0121945>
- 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>
- Evans-Hampton, T. N., Skinner, C. H., Henington, C., Sims, S., & McDaniel, C. E. (2002). An investigation of situational bias: conspicuous and covert timing during curriculum-based measurement of mathematics across african american and caucasian students. *School Psychology Review, 31*(4), 529–539. <https://doi.org/10.1080/02796015.2002.12086172>
- Furey, W. M., Marcotte, A. M., Hintze, J. M., & Shackett, C. M. (2016). Concurrent validity and classification accuracy of curriculum-based measurement for written expression. *School Psychology Quarterly, 31*(3), 369–382. <https://doi.org/10.1037/spq0000138>
- Graham, S., Hebert, M., Sandbank, M. P., & Harris, K. R. (2016). Assessing the writing achievement of young struggling writers: Application of generalizability theory. *Learning Disability Quarterly, 39*(2), 72–82. <https://doi.org/10.1177/0731948714555019>

- Han, K., Colarelli, S. M., & Weed, N. C. (2019). Methodological and statistical advances in the consideration of cultural diversity in assessment: A critical review of group classification and measurement invariance testing. *Psychological Assessment, 31*(12), 1481–1496. <https://doi.org/10.1037/pas0000731>
- Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (2002). Inferring program effects for special populations: Does special education raise achievement for students with disabilities? *Review of Economics and Statistics, 84*(4), 584–599. <https://doi.org/10.1162/003465302760556431>
- Heymans, M., & Eekhout, I. (2021). *Psfmi. Prediction model selection and performance evaluation in multiple imputed datasets* [Computer software]. <https://mwheymans.github.io/psfmi/>
- Hosp, J. L., Hosp, M. A., & Dol, J. K. (2011). Potential bias in predictive validity of universal screening measures across disaggregation subgroups. *School Psychology Review, 40*(1), 108–131. <https://doi.org/10.1080/02796015.2011.12087731>
- Kane, M. T. (2013). Validating the Interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. <https://doi.org/10.1111/jedm.12000>
- Keegan, P. J., Brown, G. T. L., & Hattie, J. A. C. (2013). A psychometric view of sociocultural factors in test validity: The development of standardised test materials for Māori medium schools in New Zealand/Aotearoa. In S. Phillipson, K. Ku, & S. N. Phillipson (Eds.), *Constructing educational achievement: A sociocultural perspective* (pp. 42–54). Routledge.
- 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, 34*(10), 2461–2480. <https://doi.org/10.1007/s11145-021-10153-6>
- Keller-Margulis, M. A., Mercer, S. H., & Thomas, E. L. (2016). Generalizability theory reliability of written expression curriculum-based measurement in universal screening. *School Psychology Quarterly, 31*(3), 383–392. <https://doi.org/10.1037/spq0000126>
- Kim, Y.-S. G., Schatschneider, C., Wanzek, J., Gatlin, B., & Al Otaiba, S. (2017). Writing evaluation: Rater and task effects on the reliability of writing scores for children in grades 3 and 4. *Reading and Writing: An Interdisciplinary Journal, 30*(6), 1287–1310. <https://doi.org/10.1007/s11145-017-9724-6>
- Kline, R. B. (2013). Assessing statistical aspects of test fairness with structural equation modelling. *Educational Research and Evaluation, 19*(2–3), 204–222. <https://doi.org/10.1080/13803611.2013.767624>
- Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 1–14). Cambridge University Press.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. <https://doi.org/10.2307/2529310>
- Licht, C. (2010). New methods for generating significance levels from multiply-imputed data [Doctoral dissertation, University of Bamberg]. Deutsch National Bibliothek. Marston, D., & Deno, S. (1981). *The Reliability of Simple, Direct Measures of Written Expression* (Vol. IRLDRR-50). University of Minnesota, Institute for Research on Learning Disabilities.
- McMaster, K. L., & Campbell, H. (2008). New and existing curriculum-based writing measures: Technical features within and across grades. *School Psychology Review, 37*(4), 550–556. <https://doi.org/10.1080/02796015.2008.12087867>
- McMaster, K. L., & Espin, C. (2007). Technical features of curriculum-based measurement in writing: A literature review. *The Journal of Special Education, 41*(2), 68–84. <https://doi.org/10.1177/00224669070410020301>
- McMaster, K. L., Shin, J., Espin, C. A., Jung, P.-G., Wayman, M. M., & Deno, S. L. (2017). Monitoring elementary students' writing progress using curriculum-based measures: Grade and gender differences. *Reading and Writing, 30*(9), 2069–2091. <https://doi.org/10.1007/s11145-017-9766-9>
- McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with coh-matrix*. 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/shmercer/writeAlizer/>

- Mercer, S. H., Keller-Margulis, M. A., Faith, E. L., Reid, E. K., & Ochs, S. The potential for automated text evaluation to improve the technical adequacy of written expression curriculum-based measurement. (2019). *Learning Disability Quarterly*, 42(2), 117–128. 803296. <https://doi.org/10.1177/0731948718>
- Muthén, L. K., & Muthén, B. O. (1998). *Mplus user's guide* (Eighth Edition ed.). Muthén & Muthén.
- Page, E. B. (2003). Project essay grade: PEG. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43–54). Erlbaum.
- Persky, H. R., Daane, M. C., & Jin, Y. (2003). *The nation's report card: Writing 2002*. (NCES 2003–529). U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics.
- 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>
- Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisaceck, F., Sanchez, J. C., & Müller, M. (2011). PROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12(1), 1–8. <https://doi.org/10.1186/1471-2105-12-77>
- Robitzsch, A., & Grund, S. (2021). *miceadds: Some additional multiple imputation functions, especially for "mice"* (R package version 3.11-6) [Computer software]. <https://CRAN.R-project.org/package=miceadds>
- 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* [Computer software]. RStudio, PBC. <http://www.rstudio.com/>
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. SAGE Publications.
- Skiba, R. J., Artiles, A. J., Kozleski, E. B., Losen, D. J., & Harry, E. G. (2016). Risks and consequences of oversimplifying educational inequities: A response to Morgan et al. *Educational Researcher*, 45(3), 221–225. 2015. <https://doi.org/10.3102/0013189X16644606>
- Skiba, R. J., Simmons, A. B., Ritter, S., Gibb, A. C., Rausch, M. K., Cuadrado, J., & Chung, C.-G. (2008). Achieving equity in special education: History, status, and current challenges. *Exceptional Children*, 74(3), 264–288. <https://doi.org/10.1177/001440290807400301>
- Smolkowski, K., Cummings, K. D., & Strycker, L. (2016). An introduction to the statistical evaluation of fluency measures with signal detection theory. In K. D. Cummings & Y. Petscher (Eds.), *The fluency construct: Curriculum-based measurement concepts and applications* (pp. 187–221). Springer. https://doi.org/10.1007/978-1-4939-2803-3_8
- Texas Education Agency (2013). *Standard setting technical report*.
- van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 1–67. <https://doi.org/10.18637/jss.v045.i03>
- Warne, R. T., Yoon, M., & Price, C. J. (2014). Exploring the various interpretations of “test bias”. *Cultural Diversity & Ethnic Minority Psychology*, 20(4), 570–582. <https://doi.org/10.1037/a0036503>
- Weissenburger, J. W., & Espin, C. A. (2005). Curriculum-based measures of writing across grade levels. *Journal of School Psychology*, 43(2), 153–169. <https://doi.org/10.1016/j.jsp.2005.03.002>
- Wilson, J., Olinghouse, N. G., McCoach, D. B., Santangelo, T., & Andrada, G. N. (2016). Comparing the accuracy of different scoring methods for identifying sixth graders at risk of failing a state writing assessment. *Assessing Writing*, 27, 11–23. <https://doi.org/10.1016/j.asw.2015.06.003>
- 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>
- Xu, Y., & Drame, E. R. (2008). Examining sociocultural factors in response to intervention models. *Childhood Education*, 85(1), 26–32. <https://doi.org/10.1080/00094056.2008.10523053>