Language Teaching Research Quarterly **EUROKD**

2025 (51), 247–272
https://doi.org/10.32038/ltrq.2025.51.04

Assessing the Accuracy of Automated Writing Evaluation in Predicting English Language Arts Proficiency for Middle-Grade English Language Learners and Non-English Language Learners

Fan Zhang
University of Delaware, USA
Central South University of Forestry and Technology, China
Joshua Wilson*
University of Delaware, USA

Correspondence
Email: joshwils@udel.edu

Abstract
This study examines the accuracy of fall, winter, and spring benchmark writing assessments, scored by the MI Write automated writing evaluation system, for predicting non-proficiency on the Smarter Balanced ELA assessment. This study considers how grade level, seasonality, and language status influence classification accuracy using Receiver Operating Characteristic (ROC) curve analyses. The results indicate that MI Write demonstrated acceptable overall classification accuracy, with the strongest performance among non-ELLs and Grade 7 students. However, accuracy was more inconsistent for ELLs. Across all grades and subgroups, the d-based cutpoints consistently provided the best balance between sensitivity and specificity. Implications for adopting AI-based assessment systems within the middle grades are discussed.

ARTICLE HISTORY
Received: 15 July 2024
Revised: 02 October 2025
Accepted: 06 November 2025

KEYWORDS
Automated Writing Evaluation (AWE), AI-Based Assessment, English Language Arts (ELA) Proficiency, Middle Grades, ROC Curve Analyses

How to cite this article (APA 7th Edition):
Zhang, F., & Wilson, J. (2025). Assessing the accuracy of automated writing evaluation in predicting English language arts proficiency for middle-grade English language learners and non-English language learners. *Language Teaching Research Quarterly*, *51*, 247–272. https://doi.org/10.32038/ltrq.2025.51.04

[1]Introduction

Writing skills are fundamental for success in school and beyond, playing a key role in academic achievement, communication, and future opportunities. Despite their importance, many middle school students in the United States (U.S.) face significant challenges in developing these skills. The 2012 National Assessment of Educational Progress (NAEP) Writing Assessment revealed that 20% of U.S. students in Grades 4 to 8 lacked the basic writing skills necessary to perform tasks at their grade level. Additionally, 54% were classified at the basic level. Notably, the percentage of eighth-grade English language learners (ELLs) who achieved proficient level or higher was only 1%, whereas 31% of non-ELLs performed at or above that level (Beck et al., 2013; National Center for Education Statistics, 2012). In the U.S., ELLs are students who come from homes where English is not the primary language and who need specialized or adapted instruction in both English and their other academic subjects (August, 2018). Writing in English is often a source of struggle for ELLs (Booth Olson et al., 2015a, 2015b). These disparities highlight the need to establish frameworks to prevent such writing difficulties and to support diverse student populations.

One such framework is Multi-Tiered Systems of Support (MTSS), which includes multi-tiered academic, social, emotional, and behavioral interventions, ensuring holistic support for all students based on their level of need (Blackburn & Witzel, 2018; Shinn et al., 2016). A central component of MTSS is universal screening, a systematic process of assessing all students within the school population to identify students who are struggling and who may need targeted intervention (Glover & Albers, 2007).

Universal screening plays a vital role in providing educational assistance based on students' needs. However, when it comes to identifying writing difficulties, universal screening is often hampered by the lack of efficient and reliable writing screeners (Berninger & Winn, 2006). To address this gap, AI-based automated writing evaluation (AWE) systems, which utilize artificial intelligence to assess writing, offer instant and consistent scoring. AWE uses advanced technologies to analyze and score longer compositions, offering consistent and reliable assessment compared to hand-scoring (Shermis, 2014; Shermis et al., 2016; Shermis & Wilson, 2024).

AWE tools have shown promise as tools for universal screening in writing (Wilson, 2018; Wilson & Rodrigues, 2020). Yet, no previous studies have specifically evaluated the classification accuracy of AWE systems with middle school students or explored their performance across different student groups, such as ELLs and non-ELLs (Wilson et al., 2024). Investigating this is important as it addresses a literature gap, ensuring that these tools can accurately identify students at risk of writing challenges, including students

[1] This paper is part of a special issue (2025, 50-51) entitled: In honour of Carol A. Chapelle's contributions to language assessment and learning (edited by Christine Coombe, Tony Clark, and Hassan Mohebbi).

from vulnerable populations such as ELLs, thereby enabling timely and appropriate interventions during this crucial educational phase.

In line with Chapelle's argument-based approach to validity for AWE (Chapelle et al., 2015), this study aims to elicit and assess validity evidence for the specific use of MI Write as a universal screening tool in middle school settings. Chapelle's framework emphasizes the need for systematic validation of automated language assessment tools by examining how well the intended use aligns with evidence from various sources, including classification accuracy across different student groups. Accordingly, the current study investigates how well the *MI Write* AWE system can predict which middle school students will not meet proficiency standards on the Smarter Balanced (SB) English language arts (ELA) assessment, a state-administered educational accountability assessment. The study considers how various factors such as the students' grade levels, the time of year the screener is administered (seasonality), and students' language status (ELLs versus non-ELL) may influence MI Write's accuracy for identifying at-risk students.

Theoretical Framework

The Simple View of Writing (SVW), proposed by Juel et al. (1986), serves as the theoretical basis for this study. The SVW emphasizes the significance of two critical skills in the development of writing: transcription and text generation. The fundamental "lower order" skills of transcription, such as handwriting, typing, and spelling, serve as a solid foundation for more sophisticated cognitive writing tasks like planning, assessing, and revising (Whitaker et al., 1994). Effective transcription skills improve students' capacity to produce and arrange ideas for written assignments (Kim et al., 2011). Transcription and text generation skills are equally important in predicting writing achievement across age groups and genres of writing (Berninger et al., 2002; Kim et al., 2013; Kim & Schatschneider, 2017). The SVW provides valuable insights for universal screening for writing, suggesting that an effective writing screener should provide information about a student's transcription and text generation skills.

Transcription fluency may be assessed through a curriculum-based measurement approach to writing assessment (W-CBM), which employs brief writing tasks, typically lasting three minutes, to evaluate writing fluency (McMaster et al., 2009). Performance on W-CBM is evaluated using metrics like Total Words Written (TWW), Words Spelled Correctly (WSC), Percentage of Words Spelled Correctly (%WSC), Correct Writing Sequences (CWS), Percentage of Correct Writing Sequences (%CWS), and Correct Minus Incorrect Writing Sequences (CIWS). However, because these W-CBM metrics focus solely on transcription fluency and the mechanical aspects of writing, they are of limited use for screening, particularly for middle schoolers. This is because they do not evaluate text generation skills, which are more often evaluated using ratings of writing quality. Developmentally, transcription is the main constraint on writing quality for younger students, whereas text generation skills, including cognitive skills needed to develop a

full text, are the primary constraint on writing quality for students in the middle grades (Berninger et al., 1991). Therefore, when screening middle school students, W-CBM, which solely measures transcription fluency, would be insufficient.

An alternative writing assessment approach, called the Direct Assessment of Writing (DAW; see Stiggins, 1982) yields inferences about students' text generation skills by assigning students longer writing tasks, often 30 minutes or more, which are then evaluated for writing quality by human raters using rubrics (Huot, 1990). Although this assessment framework is more suited to assessing the writing skills of middle schoolers, traditional human-scored DAW are not ideal for universal screening because human scoring is inefficient and subject to several sources of rater error (Englehard, 1994). AWE offers the ability to leverage the strengths of the DAW approach while mitigating its weaknesses because AWE is highly efficient and reliable (Wilson, 2018). This study thus explores AWE's accuracy as a writing screener for middle schoolers and considers whether its accuracy differs across ELLs and non-ELLs.

Literature Review

With rare exceptions (e.g., Wilson et al., 2016), prior ROC curve studies of writing screening have not focused on the middle grades, underscoring the paucity of research in this area. However, findings from prior screening studies with elementary grade students have direct relevance to the present study. For instance, Keller-Margulis et al. (2016) investigated W-CBM in 139 fourth graders, featuring 64.03% non-ELLs, 13.67% ELLs, and 22.30% Monitored (previously ELL) students. Correlations between W-CBM measures and the State of Texas Assessment of Academic Readiness (STAAR) Writing scores varied, with TWW having a moderate correlation (r = .38) and CWS showing a significant correlation (r = .42 to .50). ROC curve analysis indicated high classification accuracy (the area under the ROC curve (AUC) = .70 to .94). For ELLs, the AUC value for %CWS in the winter was .80, suggesting good screening accuracy. Limitations included small sample sizes for ELL and Monitored groups, and the focus on a single grade level.

Wilson (2018) and Wilson and Rodrigues (2020) tested the accuracy of MI Write (previously named *Project Essay Grade [PEG] Writing*) across Grades 3–5 with mixed results. Specifically, Wilson (2018) examined a sample of 230 students in Grades 3 and 4, with a demographic distribution of 51% White, 35% Hispanic or Latino, 11% Black, and 3% Asian; ELLs comprised 30% of the sample. Wilson and Rodrigues (2020) examined the accuracy of AWE as a screener across Grades 3 to 5, as well, involving 185, 167, and 187 students respectively, in a district where 36% were low-income, and over 90% of ELLs were Hispanic/Latinx and Spanish-speaking. Correlations between MI Write scores and risk status across those studies ranged from $|r|$ = .36 to .56, with AUC values from .74 to .83. Although these studies included sizable ELL samples, neither study specifically

compared the accuracy of MI Write between ELLs and non-ELLs, nor sampled middle school students.

Only one study of AWE has looked at middle school students specifically. Wilson et al. (2016) investigated the accuracy of MI Write with 272 sixth graders. MI Write scores showed moderate correlation (|*r*|= .49) and high accuracy (AUC = .78) for identifying students at risk of failing a state writing assessment. The study highlighted AWE's efficiency but noted high false positive (FP) rates.

Although AWE can be an accurate and efficient tool for writing screening, and generate reliable scores based on generalizability theory (see Chen et al., 2022; Wilson et al., 2019), limitations include high FP rates (students flagged incorrectly) (Wilson, 2018; Wilson & Rodrigues, 2020), limited research on middle school students, and potential issues with generalizability to different writing styles. Additional research applying AWE systems and adjusting their use to cater to the diverse needs of both middle school ELL and non-ELL students is needed. This will help in creating a more holistic and equitable approach to educational assessments and truly testing the viability of AWE for supporting universal screening within middle school MTSS frameworks for preventing writing difficulties.

Present Study

This study aims to address the need for effective and equitable screening methods to identify middle schoolers at risk of writing difficulties given the widespread struggle with writing in the U.S and particularly for ELL students. The present study focuses on using *MI Write*, a widely adopted AWE tool with prior promise for universal screening in the elementary grades (Wilson, 2018; Wilson et al., 2016; Wilson & Rodrigues, 2020). However, little is known about its screening accuracy in the middle grades where writing skills become more complex, nor whether accuracy of screening differs among ELLs and non-ELLs.

Drawing on Chapelle et al.'s (2015) argument-based framework for the validity of AWE, this study aims to gather and analyze systematic evidence regarding the use of AWE for universal screening purposes. Chapelle's framework emphasizes the importance of collecting multiple sources of evidence to support the intended interpretations and uses of AWE scores, especially when evaluating diverse student populations. In line with this approach, the study investigates the accuracy of MI Write in predicting non-proficiency on state assessments across different grade levels and language status, providing insights into the validity of its use as a universal screener for ELA proficiency.

Thus, two research questions guide this study:

RQ_1: How accurately does the MI Write automated writing evaluation system predict non-proficiency on the Smarter Balanced ELA assessment among middle school ELL and non-ELL students?

RQ2: Does the intersection of grade level and language status influence the classification accuracy of the MI Write automated writing evaluation system in predicting ELA proficiency among middle school ELL and non-ELL students?

Methods

Sample

The study was conducted with institutional review board approval for exempt research, and utilized data collected during the 2023–2024 school year from a mid-Atlantic school district in the U.S. The school district implemented *MI Write* for writing screening in its middle schools at three time points–Fall 2023, Winter 2024, and Spring 2024–in advance of administering the Smarter Balanced ELA (SB ELA) assessment in Spring 2024 as required by state educational accountability laws. The district educated 3,825 students in Grades 6–8 that year.

Table 1

Sample Demographics

Variable	Category	*n*	%
School	1	297	9.0
	2	413	12.5
	3	492	14.9
	4	233	7.0
	5	385	11.6
	6	690	20.9
	7	333	10.1
	8	464	14.0
School Context	Low-Needs	1892	57.2
	High-Needs	1415	42.8
Grade	6	1091	33.0
	7	1128	34.1
	8	1088	32.9
Gender	Male	1647	49.8
	Female	1660	50.2
Race/Ethnicity	White (non-Hispanic)	1357	41.0
	Black (non-Hispanic)	683	20.7
	Hispanic	1030	31.1
	Asian	214	6.5
	American Indian	18	0.5
	Pacific Islander	5	0.2
Special Education	No	2825	85.4
	Yes	482	14.6
English Language	No	2841	85.9
Learners (ELLs)	Yes	466	14.1
Honors English	No	2138	64.7
	Yes	1169	35.3

Note. N = 3,307.

The present sample was drawn from middle school students who took the Spring 2024 SB ELA assessment and completed at least one benchmark writing assessment across the academic year. This yielded a sample of 3,307 students from eight middle schools, including students in Grades 6 (n = 1,091; 33%), Grade 7 (n = 1,128; 34.1%), and Grade 8 (n = 1,088; 32.9%), with equal proportions of male students (49.8%), and female students

(50.2%). Regarding race and ethnicity, 41% of the student sample was White (non-Hispanic), 31.1% was Hispanic, 20.7% was Black (non-Hispanic), 6.5% was Asian, 0.5% was American Indian, and 0.2% was Pacific Islander. Honors students comprised 35.3% of the sample and students with disabilities comprised 14.8% of the sample. ELLs constituted 14.1% of the sample. Although additional details regarding students' primary language was not available, the district predominantly serves Spanish-speaking ELLs. Socioeconomic status, as indicated by eligibility for free or reduced-price lunch, was available at the school district level (27.33%) but unavailable for reporting at the individual student level. Table 1 presents additional demographic data on the study sample.

Measures

Smarter balanced ELA assessment

The SB ELA assessment, which served as the outcome assessment in the present study, consists of multiple-choice questions adjusted for student performance, as well as short and long writing tasks (performance tasks). Most questions are machine-scored, but some writing tasks may be hand-scored or machine-scored. The multiple-choice portion typically takes about 1.5 hours, and the writing tasks take about 2 hours (Smarter Balanced Assessment Consortium [SBAC], 2019).

Based on the item response theory model, scores on the SB ELA assessment range from 2000 to 3000. These scores correspond to four performance levels: Novice (1), Developing (2), Proficient (3), and Advanced (4). The reported overall accuracy rates for correctly classifying students into their respective performance levels were 0.81 for Grade 6, 0.82 for Grade 7, and 0.81 for Grade 8 (SBAC, 2019).

Risk status

Risk status refers to the likelihood that a student will not meet grade-level expectations in ELA. It was defined dichotomously based on the SB ELA performance level with at-risk determined by scoring below Band 3 versus not-at-risk (scoring at or above Band 3). This categorization is consistent with school accountability standards that emphasize students achieving Level 3 or higher (SBAC, 2020) and previous screening studies using state test outcome measures (Keller-Margulis et al., 2016; Wilson, 2018; Wilson & Rodrigues, 2020).

Benchmark Writing Assessments Administered and Scored by MI Write

MI Write, developed by Measurement Incorporated, was used as the writing screener, employing argumentative essay prompts based on a DAW approach. During the 2023–2024 academic year, a total of six argumentative writing prompts were administered to middle schoolers within the MI Write system, two different writing prompts per season. Within each season, the two prompts were counterbalanced across schools and

implemented at the classroom level where teachers read a scripted guide to ensure consistency.

Prompts were developed by the second author in partnership with the district ELA leadership team (see Cruz Cordero, 2024 for the full text of each prompt). Prompts were intended to be of roughly equal difficulty and did not require students to read attached source material to reduce assessment time and to ensure that inferences drawn from the writing scores were not confounded with students' reading ability. Students responded to the writing prompts based on their general background knowledge. Students were allowed up to 50 minutes (1 class period) to complete their responses.

Once students completed their responses, they submitted them to MI Write for immediate automated scoring. MI Write features a prompt-independent automated scoring model powered by the PEG scoring engine. PEG uses grade-band and genre-specific scoring algorithms to ensure consistent and accurate evaluations across various contexts. MI Write evaluates six traits of writing quality each on a 1.0 to 5.0 scale: idea development, organization, style, sentence fluency, word choice, and conventions (Coe et al., 2011; Kozlow & Bellamy, 2004). MI Write additionally offers automated feedback as well as educational tools such as interactive lessons, electronic graphic organizers, and an electronic portfolio for tracking student progress, but those features were not utilized in this application of MI Write for universal screening. Additional information about the architecture of the PEG automated scoring system and MI Write may be found in Wilson et al. (2021).

Screening/predictor variable

The main predictor variable in the present study was students' MI Write-assigned "Total Score," for the Fall (*Ftotal*), Winter (*Wtotal*), and Spring (*Wtotal*) benchmark assessments. The MI Write Total Score is formed as the sum of the scores for each of the six traits, ranging from 6.0 to 30.0. Alpha reliability for these three Total Scores ranged from α = .994–.997 across grades and language status, indicating a highly reliable aggregate measure of overall writing quality. In addition, previous research validated the reliability (Shermis, 2014) and generalizability (Wilson et al., 2019) of MI Write's Total Score.

Data Analysis

Approach for missing data

Missingness rates differed across screening periods and were non-ignorable: Fall = 19%, Winter = 18%, Spring = 34%. Little's MCAR test results showed that the data were not missing completely at random, and subsequent analyses (correlations and logistic regression) showed that missingness was systematically linked to key variables present in the dataset including school, grade, special education, honors-track-classes status, race/ethnicity, and whether the school was a high or low needs school. These factors are

historically associated with student academic performance and equity, as are well-documented within NAEP. Thus, data were treated as missing at random and we applied a form of single imputation (SI) where, instead of replacing missing values with the aggregate (grand) mean, missing values were replaced with the average of the available MI Write scores for each student. Thus, the student's own average writing performance was imputed for any missing score(s). If two scores for a student were available, then the average of the two scores was taken as the value for the missing time point. If only one score for a student was available, that score was taken as the other two timepoints. This process yielded complete data enabling analyses to be completed with the full dataset (n = 3,307).

ROC curve analysis

This study primarily relied on Receiver Operating Characteristic (ROC) curve analysis to answer its research questions. ROC curve analysis is a graph showing the performance of a classification model at all classification thresholds (Fawcett, 2006). This curve plots two parameters: True Positive Rate (students predicted to be at risk who truly are at risk) and False Positive Rate (students predicted to be at risk who are not at risk).

AUC

The primary statistic yielded by a ROC curve analysis is the *area under the ROC curve* or AUC, which is a measure of how well a screener can distinguish between two diagnostic groups (e.g., pass/fail or not-at-risk/at risk). AUC reflects the screener's accuracy, with an AUC of 1 indicating perfect accuracy, an AUC of 0.5 indicating no better than chance accuracy, and an AUC of 0 indicating complete inaccuracy. An AUC of 0.70 was used as a criterion to indicate a minimum acceptable level of screening accuracy (Hosmer et al., 2013).

Cutpoint selection

For screening models that met or exceeded the minimum acceptable AUC threshold of 0.70, a "cutpoint" on the MI Write Total Score scale was selected to optimally distinguish between students likely to pass or fail the SB ELA assessment based on whether their Total Score fell below the cutpoint (predicted to fail) or above the cutpoint (predicted to pass).

Choosing the optimal cutpoint for MI Write in predicting non-proficiency on the SB ELA assessment in middle grades involves the investigation of a series of factors (Wilson, 2018; Wilson & Rodrigues, 2020). First, we explored a d-based cutscore, which represents the point on the ROC curve closest to the upper left-hand corner—a point that represents perfect classification (Swets & Picket, 1982). This cutscore aims to maximize both sensitivity and specificity. In our analysis, sensitivity values ≥0.80 were deemed "acceptable," while values ranging from 0.70 to 0.80 were considered "borderline acceptable" (Kilgus et al., 2013; Silberglitt & Hintze, 2005).

Second, two sensitivity-based cutscores were explored. In most school-based screening studies, sensitivity is given more importance than specificity. This assumes that failing to identify at-risk students (a false negative) is more detrimental than incorrectly identifying students as at risk (a false positive) (Wilson, 2018). We explored two cutpoints, seeking to identify those at or near a sensitivity value of 80% (see Smolkowski & Cummings, 2015) and 90% (see Jenkins et al., 2007), respectively.

We then examined specificity levels for all three cutscores (i.e., the *d*-based cutscore, and the two sensitivity-based cutscores). We considered specificity values of 0.80 as "desirable," 0.70 as "acceptable," and values between 0.60 and 0.70 as "borderline acceptable" (Kilgus et al., 2013; Silberglitt & Hintze, 2005).

Additional cutpoint-specific classification metrics

To inspect the classification accuracy of the screening variables, in addition to the three cutpoints, additional threshold-dependent measures were evaluated (Wilson, 2018), including the FP rate, the False Negative (FN) rate, as well as positive predictive value (PPV), negative predictive value (NPV), and the diagnostic odds ratio (DOR) (Smolkowski & Cummings, 2015; Smolkowski et al., 2016). Catts et al. (2009) recommend a FP rate <0.50 for academic screening models. The PPV and NPV respectively represent the likelihood that a student identified as at risk or not at risk on the screener will go on to fail (PPV) or pass (NPV) the criterion measure (Wilson, 2018); PPV and NPV values of 1.00 represent perfect classification accuracy. The DOR represents the ratio of the odds of being truly at risk with the odds of being truly not at risk, and higher DOR indicates better test performance. Importantly, DOR is less dependent on the base rate of the sample than the other measures.

Power analysis

We conducted power analysis using MedCalc software (version 22.023) to ensure we had enough participants for statistically significant results. We determined the minimum sample size required to detect a meaningful difference in the AUC from a null hypothesis value of 0.50 (chance probability) (Hanley & McNeil, 1982). We set a Type I error rate at 0.05, Power of 0.80, and adopted a minimum AUC value of 0.70 (Hosmer et al., 2013). Additionally, the ratios of negative (not at risk) to positive (at risk) cases from the 2024 SB ELA data were: .704 (41.3%/58.7%) for Grade 6, .733 (42.3%/57.7%) for Grade 7, and .835 (45.5%/54.5%) for Grade 8. The power analysis showed that the study would be adequately powered with minimum sample sizes of 64 Grade 6 students (27 negative and 37 positive cases), 63 Grade 7 students (27 negative and 36 positive cases), and 63 Grade 8 students (29 negative and 34 positive cases). Those sample sizes were all exceeded indicating adequate power for grade-level analyses.

The ratios of negative to positive cases were .064 (6%/94%) for ELLs and 0.965 (49.1%/50.9%) for non-ELLs. The power analysis indicated that the study would be

adequately powered with minimum sample sizes of 63 non-ELLs (31 negative and 32 positive cases) and 254 ELLs (16 negative and 238 positive) for ELLs. Our study was thus adequately powered for both aggregate and grade-specific analyses for the non-ELL subsample and for the aggregate sample of ELLs across grades, but not for grade-specific analyses of ELLs (see Table 2 for sample sizes). Thus, caution should be taken when interpreting the grade-specific analyses for the ELL subsample.

Results

Descriptives and Correlations

Descriptive statistics for assessment scores by language status and grade level are presented in Table 2. With one exception (Grade 8 non-ELLs), MI Write Total Scores tended to increase from Fall to Spring within a grade level for both non-ELLs and ELLs. Across grades and seasons, Grade 8 students tended to score the highest, followed by Grade 6 students; Grade 7 tended to score the lowest, but that was likely an artifact of the MI Write scoring system that uses grade-band specific scoring algorithms: students in Grades 5–6 are scored in a different grade band than students in Grades 7–8. Students in the lower part of a grade band (e.g., Grade 7 students) will score lower than grades in the upper part of an adjacent grade band (e.g., Grade 6 students).

Table 2

Descriptive Statistics by Language Status and Grade Level

Language Status	Assessment	Statistic	Grade 6	Grade 7	Grade 8
Non-ELL	Ftotal	*M*	16.30	14.50	16.76
		SD	5.05	4.83	4.61
	Wtotal	*M*	16.46	15.23	17.21
		SD	5.04	5.08	4.82
	Stotal	*M*	17.29	15.72	16.84
		SD	5.31	5.15	4.78
	SB24score	*M*	2519.93	2539.27	2567.12
		SD	103.82	109.29	104.606
			n = 943	*n* = 977	*n* = 915
ELL	Ftotal	*M*	12.80	11.11	13.05
		SD	4.53	3.47	3.95
	Wtotal	*M*	12.81	11.52	13.28
		SD	4.63	4.07	4.11
	Stotal	*M*	13.45	12.24	13.61
		SD	4.67	4.23	4.22
	SB24score	*M*	2409.39	2434.78	2454.68
		SD	67.86	80.707	82.73
			n = 147	*n* = 150	*n* = 167

Note. Ftotal = Fall Total Score measured by MI Write (range = 6–30); Wtotal = Winter Total Score measured by MI Write (range = 6–30); Stotal = Spring total score measured by MI Write (range = 6–30); SB24score = Smarter Balanced 2024 scale score (range = 2000–3000).

Correlations between MI Write Total Scores and the SB ELA score by language status and grade level are presented in Table 3. Predictor–Criterion relationships ranged from *r* = .508 to.659 for non-ELLs and from *r* = .296 to .446 for ELLs. There was no clear pattern

with respect to seasonality. For Grades 6 and 7, predictor–criterion relationships were strongest in the spring. Predictor criterion relationships tended to be weaker in the winter. Although, for Grade 8 non-ELLs, the relationship was strongest in winter, but for Grade 8 ELLs, it was strongest in fall and decreased each season.

Table 3

Correlation Matrix for Grades 6, 7, and 8 by Language Status (Non-ELL and ELL)

Grade and Status	Variable	Ftotal	Wtotal	Stotal	SB24score
Grade 6 Non-ELL	Ftotal	-			
	Wtotal	.753	-		
	Stotal	.770	.768	-	
	SB24score	.601	.607	.602	-
Grade 6 ELL	Ftotal	-			
	Wtotal	.704	-		
	Stotal	.749	.757	-	
	SB24score	.392	.365	.416	-
Grade 7 Non-ELL	Ftotal	-			
	Wtotal	.784	-		
	Stotal	.796	.827	-	
	SB24score	.613	.635	.659	-
Grade 7 ELL	Ftotal	-			
	Wtotal	.684	-		
	Stotal	.697	.740	-	
	SB24score	.429	.296	.446	-
Grade 8 Non-ELL	Ftotal	-			
	Wtotal	.745	-		
	Stotal	.633	.697	-	
	SB24score	.517	.558	.508	-
Grade 8 ELL	Ftotal	-			
	Wtotal	.850	-		
	Stotal	.857	.875	-	
	SB24score	.431	.428	.396	-

Note. Ftotal = Fall Total Score measured by MI Write (range = 6–30); Wtotal = Winter Total Score measured by MI Write (range = 6–30); Stotal = Spring total score measured by MI Write (range = 6–30); SB24score = Smarter Balanced 2024 scale score (range = 2000–3000). All correlations are statistically significant at $p < .01$.

RQ1: Overall Classification Accuracy

Full sample

Table 4 reports the AUC values, *d*-based and sensitivity-based cutpoints, along with their respective threshold-dependent measures of classification accuracy. As shown in that table, the AUC values suggest that MI Write performed well as a screener in the full

sample, generally providing good discriminatory power between those at risk and not at risk. AUC values ranged from .792 in the fall to approximately .800 in both the winter and spring.

In all cases, the 90% sensitivity cutpoint did not yield acceptable specificity, with values ranging from .40 (fall) to .43 (winter) and corresponding FP Rates at or approaching .60 in all three seasons. This cutpoint also yielded the lowest DOR values. The 80% sensitivity cutpoint yielded borderline acceptable sensitivity, borderline acceptable specificity, and FP rates, with a range of sensitivity between .62–.64. This cutpoint yielded the second highest DOR values. Although the *d*-based cutpoint also provided sensitivity values falling in the borderline acceptable range, specificity values were much higher (range = .72–.76) than both the 80%- and 90%-sensitivity cutpoints, falling in the acceptable range. Thus, because the *d*-based cutpoints yielded more acceptable FP rates and the highest DOR values, these *d*-based cutpoints of the MI Write Overall Score should be selected to achieve maximal classification accuracy for the full sample.

In conclusion, MI Write displayed good AUCs for all three time points. Notably, at least one cutpoint (i.e., *d*-based one) met acceptable classification accuracy criteria across all three time points, indicating that MI Write is promising in identifying students who are at risk and who are not in passing the SB ELA test.

Non-ELLs

Table 4 presents the AUC values, *d*-based and sensitivity-based cutpoints, along with their respective measures of classification accuracy for the non-ELL subgroup. The AUC values indicate that MI Write maintains good discriminatory power within this group, with AUC values ranging from .780 in the fall to approximately .790 in the winter and spring.

Similar to the full sample, the 90% sensitivity cutpoint for non-ELLs did not yield acceptable specificity, with specificity values ranging from .38 (fall) to .40 (winter and spring), resulting in FP rates approaching .62. This cutpoint also recorded the lowest DOR values within this subgroup. The 80% sensitivity cutpoint achieved borderline acceptable specificity, ranging from .57 to .61. The *d*-based cutpoint showed improved sensitivity, increasing from .67 in the fall to .70 in the winter and .72 in the spring. Specificity values were also higher (range = .73–.76). Therefore, the *d*-based cutpoints for the MI Write Overall Score should be selected to achieve maximal classification accuracy for the non-ELL subgroup.

In conclusion, given the results, MI Write also showed promising discriminatory power between those at risk and not at risk for the non-ELL subgroup.

ELLs

Table 4 illustrates the AUC values, *d*-based and sensitivity-based cutpoints, along with their respective measures of classification accuracy for the ELL subgroup. The AUC values for this subgroup were lower than those observed for non-ELLs (.729 in the fall and .702 in the spring), which indicates that MI Write has comparatively weaker discriminatory power in identifying at-risk students within the ELL subgroup. Notably, the winter assessment recorded an AUC value less than .70 (i.e., .678), suggesting that MI Write had limited discriminatory effectiveness during this period.

For the ELL subgroup, the 90% sensitivity cutpoint resulted in very low specificity values ranging from .29 in the winter to .39 in the spring, leading to high FP rates of .71 in the fall and .61 in the spring. The 80% sensitivity cutpoint, while providing slightly better specificity (.54 for fall and .43 for spring), still produced relatively high FP rates of .46 in the fall and .57 in the spring respectively. In contrast, the *d*-based cutpoint maintained moderate sensitivity values of .70 in both the fall and the spring, with corresponding specificity values ranging from .64 to .68. Although the *d*-based cutpoint resulted in borderline acceptable sensitivity and specificity, FP rates were more acceptable than the sensitivity-based cutpoints, falling within the recommended criterion of <.50 (i.e., .32 in the fall and .36 in the spring) and thus providing a more balanced trade-off between sensitivity and specificity.

Overall, MI Write showed modest discriminatory power within the ELL subgroup, particularly in the fall and spring. However, its effectiveness was notably reduced during the winter season.

Table 4

AUCs, Cutpoints, and Classification Accuracy Statistics for the Full Sample, All Non-ELLs, and All ELLs

	Assessment	AUC (SE)	95% CI	Cutpoint	Sensitivity	Specificity	False Positive Rate	False Negative Rate	Positive Predictive Value	Negative Predictive Value	Diagnostic Odds Ratio
Full Sample	Ftotal	0.792	[0.777,	15.350	0.70	0.76	0.24	0.30	0.79	0.65	7.33
		(0.008)	0.807]	17.150	0.80	0.62	0.38	0.20	0.73	0.70	6.44
				19.450	0.90	0.40	0.60	0.10	0.67	0.75	5.99
	Wtotal	0.804	[0.789,	16.275	0.73	0.74	0.26	0.27	0.79	0.67	7.42
		(0.008)	0.818]	17.475	0.80	0.64	0.36	0.20	0.75	0.71	7.09
				19.850	0.90	0.43	0.57	0.10	0.68	0.77	6.87
	Stotal	0.796	[0.781,	16.925	0.74	0.72	0.28	0.26	0.78	0.68	7.62
		(0.008)	0.811]	18.075	0.80	0.63	0.37	0.20	0.74	0.70	6.83
				20.325	0.90	0.42	0.58	0.10	0.67	0.76	6.52
Non-ELLs	Ftotal	0.778	[0.761,	15.450	0.67	0.76	0.24	0.33	0.74	0.69	6.40
		(0.009)	0.795]	17.725	0.80	0.57	0.43	0.20	0.66	0.73	5.35
				19.725	0.90	0.38	0.62	0.10	0.60	0.78	5.46
	Wtotal	0.791	[0.775,	16.375	0.70	0.74	0.26	0.30	0.73	0.70	6.53
		(0.008)	0.808]	17.825	0.80	0.61	0.39	0.20	0.68	0.75	6.35
				20.350	0.90	0.40	0.60	0.10	0.61	0.80	6.05
	Stotal	0.785	[0.769,	16.925	0.72	0.73	0.27	0.28	0.73	0.72	6.96
		(0.008)	0.802]	18.375	0.80	0.61	0.39	0.20	0.68	0.74	6.22
				20.550	0.90	0.40	0.60	0.10	0.61	0.79	5.85
ELLs	Ftotal	0.729	[0.642,	14.050	0.70	0.68	0.32	0.30	0.97	0.13	4.89
		(0.044)	0.816]	15.550	0.80	0.54	0.46	0.20	0.96	0.15	4.72
				17.700	0.90	0.29	0.71	0.10	0.95	0.15	3.58
	Wtotal	0.678	[0.570,								
		(0.055)	0.786]								
	Stotal	0.702	[0.600,	14.750	0.70	0.64	0.36	0.30	0.97	0.12	4.17
		(0.052)	0.803]	16.300	0.80	0.43	0.57	0.20	0.96	0.12	2.98
				18.350	0.88	0.39	0.61	0.12	0.96	0.17	4.70

Note. Ftotal, Wtotal, and Stotal represent the MI Write Total Score (range = 6–30) for fall, winter, and spring assessment administrations, respectively. For the columns presenting the cutpoints and the associated threshold-dependent classification accuracy metrics, the three values arranged vertically correspond to the *d*-based cutpoint, 80%-sensitivity cutpoint, and the 90%-sensitivity cutpoint, respectively, in order from top to bottom within the cell.

Table 5

*AUCs, Cutpoints, and Classification Accuracy Statistics for the Intersection of Season*Grade*Language*

	Assessment	AUC (SE)	95% CI	Cutpoint	Sensitivity	Specificity	False Positive Rate	False Negative Rate	Positive Predictive Value	Negative Predictive Value	Diagnostic Odds Ratio
Grade 6 non-ELLs	Ftotal	0.784 (0.015)	[0.755, 0.813]	16.950	0.74	0.68	0.32	0.26	0.72	0.71	6.26
				18.075	0.80	0.61	0.39	0.21	0.69	0.73	5.96
				20.150	0.90	0.43	0.57	0.10	0.64	0.80	6.86
	Wtotal	0.790 (0.015)	[0.761, 0.819]	15.650	0.65	0.80	0.20	0.35	0.79	0.67	7.51
				18.075	0.80	0.61	0.39	0.20	0.69	0.73	6.04
				20.750	0.90	0.38	0.62	0.10	0.62	0.77	5.51
	Stotal	0.790 (0.015)	[0.762, 0.819]	17.725	0.73	0.73	0.27	0.27	0.75	0.71	7.22
				19.450	0.80	0.59	0.41	0.20	0.69	0.73	5.78
				21.850	0.90	0.38	0.62	0.10	0.62	0.77	5.38
Grade 6 ELLs	Ftotal	0.728 (0.074)	[0.583, 0.874]	14.700	0.69	0.75	0.25	0.31	0.99	0.06	6.53
				16.350	0.80	0.50	0.50	0.20	0.98	0.06	3.93
				16.850	0.83	0.50	0.50	0.17	0.98	0.07	4.72
	Wtotal	0.730 (0.106)	[0.522, 0.937]	15.775	0.78	0.75	0.25	0.22	0.99	0.09	10.41
				16.050	0.78	0.50	0.50	0.22	0.98	0.06	3.61
				21.300	0.95	0.25	0.75	0.05	0.98	0.13	6.48
	Stotal	0.691 (0.159)	[0.379, 1.000]	-	-	-	-	-	-	-	-
Grade 7 non-ELLs	Ftotal	0.806 (0.014)	[0.779, 0.833]	14.350	0.71	0.76	0.24	0.29	0.76	0.70	7.68
				15.450	0.80	0.68	0.32	0.21	0.73	0.75	8.17
				18.275	0.90	0.40	0.60	0.10	0.62	0.78	6.00
	Wtotal	0.813 (0.013)	[0.787, 0.840]	14.950	0.71	0.76	0.24	0.29	0.76	0.71	7.74
				16.625	0.80	0.64	0.36	0.20	0.71	0.74	6.89
				18.550	0.90	0.47	0.53	0.10	0.65	0.81	7.95
	Stotal	0.825 (0.013)	[0.800, 0.851]	15.950	0.76	0.75	0.25	0.24	0.77	0.74	9.55
				16.875	0.80	0.69	0.31	0.20	0.74	0.76	8.97
				19.150	0.90	0.52	0.48	0.10	0.67	0.82	9.60

	Assessment	AUC (SE)	95% CI	Cutpoint	Sensitivity	Specificity	False Positive Rate	False Negative Rate	Positive Predictive Value	Negative Predictive Value	Diagnostic Odds Ratio
Grade 7 ELLs	Ftotal	0.791 (0.075)	[0.644, 0.939]	12.800	0.72	0.82	0.18	0.28	0.98	0.19	11.54
				13.800	0.81	0.64	0.36	0.19	0.97	0.21	7.61
				15.850	0.94	0.55	0.45	0.06	0.96	0.40	17.33
	Wtotal	0.621 (0.097)	[0.432, 0.811]	-	-	-	-	-	-	-	-
	Stotal	0.713 (0.081)	[0.553, 0.872]	14.275	0.73	0.64	0.36	0.27	0.96	0.16	4.65
				15.250	0.80	0.55	0.45	0.20	0.96	0.18	4.76
				17.800	0.91	0.36	0.64	0.09	0.95	0.24	5.54
Grade 8 non-ELLs	Ftotal	0.753 (0.016)	[0.721, 0.784]	16.950	0.68	0.72	0.28	0.32	0.68	0.71	5.34
				18.550	0.80	0.55	0.45	0.21	0.61	0.75	4.63
				20.250	0.90	0.36	0.64	0.10	0.56	0.80	5.05
	Wtotal	0.773 (0.015)	[0.743, 0.803]	17.150	0.69	0.75	0.25	0.31	0.71	0.73	6.68
				19.250	0.80	0.55	0.45	0.20	0.62	0.75	4.80
				21.025	0.90	0.36	0.64	0.10	0.56	0.80	5.22
	Stotal	0.743 (0.016)	[0.711, 0.775]	16.925	0.65	0.73	0.27	0.35	0.68	0.70	5.02
				18.950	0.80	0.54	0.46	0.20	0.61	0.75	4.63
				20.550	0.90	0.36	0.64	0.10	0.56	0.80	5.13
Grade 8 ELLs	Ftotal	0.698 (0.072)	[0.556, 0.840]	-	-	-	-	-	-	-	-
	Wtotal	0.724 (0.077)	[0.574, 0.874]	14.850	0.71	0.69	0.31	0.29	0.96	0.16	5.38
				16.150	0.80	0.62	0.38	0.21	0.96	0.20	6.20
				20.000	0.94	0.31	0.69	0.06	0.94	0.31	7.26
	Stotal	0.712 (0.077)	[0.562, 0.862]	14.700	0.68	0.69	0.31	0.32	0.96	0.15	4.77
				17.250	0.80	0.46	0.54	0.20	0.95	0.16	3.46
				19.750	0.92	0.38	0.62	0.08	0.95	0.29	7.50

Note. Ftotal, Wtotal, and Stotal represent the MI Write Total Score (range = 6–30) for fall, winter, and spring assessment administrations, respectively. For the columns presenting the cutpoints and the associated threshold-dependent classification accuracy metrics, the three values arranged vertically correspond to the *d*-based cutpoint, 80%-sensitivity cutpoint, and the 90%-sensitivity cutpoint, respectively, in order from top to bottom within the cell.

RQ2: Accuracy by Grade Level and Language Status

Grade 6

As shown in Table 5, for non-ELLs in Grade 6, across all assessment periods, MI Write showed a consistent ability to identify at-risk and not-at-risk students, with all AUC values exceeding .75. The 90% sensitivity cutpoint resulted in very low specificity and high FP rates across the three seasons. For example, in the fall, the 90% sensitivity-based cutpoint was .90 with specificity at .43 and a high FP rate of .57, with other seasons showing a similar pattern. Conversely, the 80% sensitivity cutpoint showed improved balance; in the fall, sensitivity at the *d*-based 80% cutpoint was .80, specificity was .61, and the FP rate was .39, consistently below .50 across all seasons. While the 80% sensitivity cutpoint presented a better balance compared to the 90% sensitivity cutpoint, the *d*-based cutpoint generally offered the best balance, achieving moderate sensitivity and specificity in most seasons, except for the winter. During the winter, although specificity was desirable, sensitivity dropped to .65, falling below the borderline acceptability threshold. The DOR values for the seasons under the *d*-based cutpoints also indicated stronger predictive capabilities, with values such as 7.51 in the winter and 7.22 in the spring. Therefore, *d*-based cutpoints should be selected to optimize classification accuracy.

Grade 6 ELLs showed generally lower AUC values compared to non-ELLs, and MI Write's effectiveness in detecting at-risk ELL students in Grade 6 notably declined during the spring season. The pattern indicated more challenges in achieving both desirable sensitivity and specificity. High sensitivity levels (90% and 80%) were often associated with very low specificity and correspondingly high FP rates. In contrast, the *d*-based cutpoint provided a better balance between sensitivity and specificity. In the fall, it achieved nearly borderline acceptable sensitivity of .69, with a specificity of .75, and a FP rate of .25. In the winter, sensitivity improved to .78 with the same specificity and FP rate. The DOR values were 6.53 and 10.41 for the two seasons, respectively.

Grade 7

For Grade 7 non-ELLs, AUC values ranged from .806 to .825, indicating good accuracy in distinguishing between students who are at risk and those who are not. At the 90% sensitivity-based cutpoints, specificity was low (ranging from .40 to .52), and the FP rates were relatively high, ranging from .48 to .60. DOR ranged from 7.68 to 9.55, suggesting a good diagnostic ability of the tool. The 80% sensitivity-based cutpoints demonstrated borderline specificity (ranging from 0.64 to 0.69), acceptable FP rates (between .31 and .36), and DOR values from 6.89 to 8.97. Similarly, the *d*-based cutpoints offered a better balance between sensitivity and specificity, along with strong DOR values, making them the optimal choice for classification cutpoints for this grade level and subgroup.

The AUC values for Grade 7 ELLs are generally lower compared to non-ELLs, ranging from .621 to .791, indicating lesser predictive capability. Notably, MI Write's ability to identify

at-risk ELL students in Grade 7 significantly decreased during the winter season, falling below the minimal acceptable AUC value of 0.70. The 90% sensitivity cutpoint was high but at the cost of specificity (.55 and .36 in fall and spring, respectively), and the FP rates for these cutpoints were correspondingly high at .45 and .64. DOR values for Grade 7 ELLs varied significantly for the two seasons (range = 5.54–17.33) and more so than for non-ELLs (range = 6.00–9.60). The 80% sensitivity cutpoints presented similar patterns as the 90% cutpoints. The *d*-based cutpoints were also the optimal choice for classification cutpoints.

Grade 8

As illustrated in Table 5 for Grade 8 non-ELLs and ELLs, MI Write displayed varying levels of screening accuracy. For non-ELLs, AUC values ranged from .743 to .773, suggesting good predictive accuracy. Across different assessment periods, the 90% sensitivity cutpoint showed high sensitivity (.90) but low specificity (.36), resulting in high FP rates (.64). The 80% sensitivity cutpoint provided high specificity (.55) with lower FP rates (.45). The DOR values for the two sensitivity-based cutpoints ranged from 4.63 to 5.22. Although the *d*-based cutpoints provided unacceptable sensitivity (between .65 and .69), they yielded acceptable specificity (.72–.75) and stronger DOR values (between 5.02 and 6.68), indicating greater discriminatory effectiveness for identifying not-at-risk students.

AUC Values for Grade 8 ELLs were generally lower than those for non-ELLs, ranging from .698 to .724. During the fall season, MI Write's effectiveness notably declined, falling below the .70 threshold for acceptable AUC values. Cutpoint for Grade 8 ELLs consistently showed lower specificity and higher FP rates. For example, sensitivity at the 90% cutpoint in the winter was very high (.94) but with low specificity (.31), leading to a very high FP rate (.69). The DOR values across the winter and spring ranged from 3.46 to 6.20. In the winter, the sensitivity at the 80% cutpoint was matched with a specificity of .62 and a FP rate of .38. The DOR values ranged from 7.26 to 7.50. However, the *d*-based cutpoints offered a better balance of sensitivity and specificity compared to the 90% and 80% sensitivity-based cutpoints.

Discussion

ELL students in U.S. middle schools often struggle with writing, and there is limited research on the accuracy of AWE tools in identifying at-risk and not-at-risk students in this demographic group. This study aims to address this gap by investigating effective and equitable screening methods for writing challenges among middle school students. Specifically, it examined the accuracy of MI Write as a universal writing screener in predicting performance on the SB ELA assessments among ELL and non-ELL students in middle grades, where writing skills become increasingly complex. Specifically, the study evaluated the accuracy of benchmark assessments scored by MI Write across three seasons in predicting non-proficiency on the SB ELA assessment. It also explored whether

grade level and language status impact the classification accuracy of MI Write in predicting ELA proficiency among middle school ELL and non-ELL students.

When not-disaggregating by grade level (RQ1), the ROC curve analyses revealed that MI Write displayed overall acceptable classification accuracy in identifying at-risk and not-at-risk students for the SB ELA tests, with all nine analyses meeting the 0.70 AUC threshold in the combined sample of ELL and non-ELL students in Grades 6-8. For non-ELLs, MI Write consistently achieved AUC values above 0.70 across all grades and seasons, indicating its promise in distinguishing students with or without writing difficulties. However, the same level of accuracy was not observed among ELLs, who exhibited more variability and lower accuracy in MI Write's screening effectiveness (range AUC = .678–.729). In all cases, *d*-based cutpoints proved superior.

Moving to RQ2, across all grades and seasons, MI Write was more effective for screening non-ELLs, as there were consistently higher AUC values for this language subgroup, indicating a robust ability of MI Write to discriminate effectively between at-risk and not-at-risk students. In addition, it was possible to identify a *d*-based cutpoint that yielded acceptable or borderline acceptable sensitivity and specificity values for non-ELLs, as well as higher DOR values. However, MI Write exhibited lower AUC values when applied to ELLs across all grades and seasons, especially in certain seasons, nor was it possible to consistently identify a cutpoint that yielded acceptable or borderline acceptable sensitivity, specificity, and FP rates. Moreover, the DOR values for ELLs were generally lower compared to non-ELLs, reflecting the reduced screening effectiveness of MI Write among the ELL group.

The findings of this study align with prior research, such as that of Keller-Margulis et al. (2016), which emphasized the difficulty of identifying writing screeners and cutpoints that achieve a satisfactory balance of sensitivity and specificity for screening ELLs. Their study also found higher AUC values for non-ELLs (referred to as "Native Speakers") and identified several W-CBM metrics with cutscores meeting criteria for acceptable classification accuracy. The challenge is further highlighted by the fact that Keller-Margulis et al. (2016) used writing screeners to predict performance on a state writing test, a context where one might expect greater prediction accuracy compared to the present study, which used writing screeners to predict outcomes on a state English language arts assessment encompassing both reading and writing.

Findings are also consistent with prior studies of MI Write by Wilson and colleagues (Wilson, 2018; Wilson et al., 2016; Wilson & Rodrigues, 2020), whose research indicated that MI Write shows promise for universal screening when used in aggregate samples (i.e., non-ELLs and ELLs) in elementary and middle grades. Like these prior studies, the present study also reported AUC values in the .70–.80 range, suggesting acceptable

classification accuracy but not sufficiently high for making high-stakes decisions without further evaluation.

These results underscore the importance of evaluating the validity of AWE systems for specific subgroups, an approach that aligns with Chapelle's argument-based validity framework for AWE (Chapelle et al., 2015). Chapelle's framework emphasizes collecting systematic evidence to support the intended interpretations and uses of test scores, particularly for diverse populations. In the context of this study, the lower accuracy of MI Write for ELLs suggests that further validation efforts are needed to ensure that the tool provides equitable screening across student groups.

Building on this point, the current study suggests that while AWE shows promise for writing screening, its implementation should be approached with caution. One viable approach may be to use AWE as a universal screener for the entire middle school population, especially for non-ELLs, while incorporating supplementary evaluation measures, such as human-scored writing assessments evaluating various levels of language such as spelling, vocabulary, sentence structure and fluency, and discourse-level writing skills (e.g., organization and idea development), for ELLs flagged as at-risk during the initial screening. Indeed, prior research by Chen et al. (2022) compared the reliability of the MI Write Total Score with human ratings of writing quality for evaluating the writing of struggling and non-struggling writers in Grades 3–5. Results of multivariate generalizability theory analyses reported greater reliability when evaluating non-struggling writers than struggling writers. The authors recommended that MI Write be adopted for use for periodic classwide formative assessment and to conduct follow-up human-scored writing assessments to support accurate inferences regarding struggling writers. Our recommendation is thus consistent with prior research, suggesting that while AWE systems like MI Write can be effective for large-scale formative assessment, particularly with non-struggling writers, additional human-scored assessments may be necessary to ensure reliable and valid evaluation of struggling writers, including ELLs.

Future research should explore a two-staged gated screening approach to consider the potential benefits of using AWE as an initial screener, followed by more targeted, resource-intensive assessments for students flagged as at-risk. AWE could be used to rule out students who are not-at-risk, and follow-up human-scored writing assessments can be used to reduce FP rates among the students identified as at risk. This approach could help balance the need for broad coverage and efficiency with the necessity for more accurate identification of students who may require additional support, especially among ELLs. By applying a more sensitive follow-up assessment only to those initially identified, a gated approach may improve overall classification accuracy, reduce the burden of extensive human scoring, and better allocate educational resources. Future studies should examine whether the follow-up assessments should be in the domain of writing, reading, or ELL-specific language assessments. Indeed, prior research on gated screening

suggests that follow-up assessments offer different levels of added value in gated screening models (Van Norman et al., 2019).

Limitations and Future Directions

Results should be interpreted relative to the following study limitations. First, the dependent variable, risk status, is derived from performance on the SB ELA assessment, which evaluates both reading and writing skills. The screening models might have demonstrated stronger results if the dependent variable had been exclusively focused on writing performance or had we tested the accuracy of a combined screening model, wherein we administered both reading and writing assessments, and tested the combination of those assessments for predicting SB ELA performance. Future research should test a combined screening model that incorporates both reading and writing assessments to determine whether this integrated approach improves the prediction of SB ELA performance. This could help clarify whether combining multiple indicators enhances classification accuracy by capturing a broader range of language skills, thereby providing a more comprehensive evaluation of students' overall English language arts proficiency. Additionally, this approach should be investigated using a gated screening procedure, where only students flagged by the initial writing assessment undergo further reading evaluation, as opposed to administering both assessments to all students. Conversely, screening all students with a reading assessment first, followed by a targeted writing assessment for those identified as at-risk, is another option worth exploring to determine the most efficient and accurate screening sequence.

Second, factors such as individual differences in student motivation and school context might have influenced study findings. The district did not place any stakes on the benchmark writing assessments, which may have reduced student effort and performance and thereby increasing the likelihood of false positives, particularly for capable students who might lack motivation due to the lack of consequences. In addition, although we counterbalanced two writing prompts across schools each administration season, these prompts were not subjected to an equating process and thus may vary in difficulty, potentially influencing the accuracy of screening models. Future research should explore the performance of AWE-based screening models in contexts where higher stakes are attached to the assessment outcomes, potentially increasing student motivation and effort. This could help determine whether a more motivated testing environment reduces the likelihood of false positives and enhances the accuracy of the screening process.

Finally, although grade-specific sample sizes for ELLs were larger compared to previous studies that deliberately sampled ELLs (e.g., Keller-Margulis et al., 2016) or included ELLs in an aggregate sample (Wilson, 2018; Wilson & Rodrigues, 2020), these grade-specific ELL sample sizes were relatively small, which may affect the generalizability of the findings. Indeed, the study lacked power for the grade-specific analyses of the ELL

subsample. Thus, caution should be taken when interpreting those analyses. Relatedly, the ELL samples demonstrated restricted range of performance on the MI Write total score and nearly 90% of ELLs were in the "at risk" category based on their performance on the 2024 SBELA assessment. These limitations hinder the development of an accurate screening model, as variability in data is crucial for robust statistical analysis. These limitations highlight the need for future research to consider these factors for enhancing the validity and reliability of screening assessments in educational settings.

Conclusion

While the findings suggest that MI Write can effectively identify at-risk students for English language arts challenges, its accuracy is more limited when applied to ELL populations. The results underscore the need for practitioners and policymakers to take a cautious approach that combines AWE with additional assessment methods, especially for subgroups where screening effectiveness is reduced. A gated screening model, where AWE serves as an initial filter followed by more targeted assessments, offers a promising strategy to improve accuracy while minimizing resource demands. AWE developers may wish to consider evaluating and updating their scoring algorithms to ensure that the training data reflects the diversity of the student populations the system aims to serve. For example, future research could explore the classification accuracy of scoring models trained on datasets with varying proportions of ELLs, helping to determine whether a more balanced representation in training data improves the system's performance across different subgroups. In sum, the present study offers important insights for ensuring that AI-based AWE systems like MI Write serve as effective and equitable educational assessment tools across diverse student populations. To move from research to practice, school and district leaders could integrate gated screening models into their MTSS frameworks, ensuring strategic use of tools like MI Write. This process should be supported by clear school- and district-level policy guidance, staff training, and ongoing evaluation to enhance early identification efforts while addressing the unique needs of diverse student subgroups.

ORCID

https://orcid.org/0000-0001-6201-9143
https://orcid.org/0000-0002-7192-3510

Publisher's Note

The claims, arguments, and counter-arguments made in this article are exclusively those of the contributing authors. Hence, they do not necessarily represent the viewpoints of the authors' affiliated institutions, or EUROKD as the publisher, the editors and the reviewers of the article.

Acknowledgements

Not applicable.

Funding
The authors received no specific funding for this work.

CRediT Authorship Contribution Statement
Fan Zhang: Conceptualization, Methodology, Formal Analysis, Investigation, Writing – Original Draft, Writing – Review & EditingJoshua Wilson: Conceptualization, Methodology, Formal Analysis, Investigation, Writing – Original Draft, Writing – Review & Editing, Supervision

Generative AI Use Disclosure Statement
The authors utilized ChatGPT 4 to revise portions of this text for clarity.

Ethics Declarations
World Medical Association (WMA) Declaration of Helsinki–Ethical Principles for Medical Research Involving Human Participants
All procedures performed in this educational study adhered to the general ethical principles for research involving human participants, as outlined in the World Medical Association (WMA) Declaration of Helsinki (2013). The study protocol was specifically reviewed and approved by the University of Delaware Institutional Review Board (IRB) under the Exempt category, and the IRB granted a Waiver of Documentation of Consent.

Competing Interests
The authors declear no competing interests relative to this work.

Data Availability
Data is available upon reasonable request to the corresponding author.

References

August, D. (2018). Educating English language learners: A review of the latest research. *American Educator*, *42*(3), 4. Downloaded from https://files.eric.ed.gov/fulltext/EJ1192670.pdf

Beck, I. L., McKeown, M. G., & Kucan, L. (2013). *Bringing words to life: Robust vocabulary instruction.* Guilford Press.

Berninger, V. W., Abbott, R. D., Abbott, S. P., Graham, S., & Richards, T. (2002). Writing and reading: Connections between language by hand and language by eye. *Journal of learning disabilities*, 35(1), 39-56. https://doi.org/10.1177/002221940203500104

Berninger, V. W., Mizokawa, D. T., & Bragg, R. (1991). Theory-based diagnosis and remediation of writing disabilities. *Journal of School Psychology*, 29, 57–79. https://doi.org/10.1016/0022-4405(91)90016-K

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). Guilford Press.

Blackburn, B. R., & Witzel, B. S. (2018). *Rigor in the RTI and MTSS classroom: Practical tools and strategies*. Routledge. https://doi.org/10.4324/9781315639406

Booth Olson, C., Scarcella, R., & Matuchniak, T. (2015a). English learners, writing, and the Common Core. *The Elementary School Journal, 115*(4), 570–592. https://doi.org/10.1086/681235

Booth Olson, C., Scarcella, R. C., & Matuchniak, T. (2015b). *Helping English learners to write: Meeting Common Core standards, grades 6–12*. Teachers College Press.

Catts, H. W., Petscher, Y., Schatschneider, C., Bridges, M. S., & Mendoza, K. (2009). Floor effects associated with universal screening and their impact on early identification of reading disabilities. *Journal of Learning Disabilities*, 42, 163–176. https://doi.org/10.1177/0022219408326219

Chapelle, C. A., Cotos, E., & Lee, J. (2015). Validity arguments for diagnostic assessment using automated writing evaluation. *Language Testing, 32*(3), 385–405. https://doi.org/10.1177/02655322145653

Chen, D., Hebert, M., & Wilson, J. (2022). Examining human and automated ratings of elementary students' writing quality: A multivariate generalizability theory application. *American Educational Research Journal, 59*(6), 1122–1156. https://doi.org/10.3102/00028312221106773

Coe, M., Hanita, M., Nishioka, V., & Smiley, R. (2011). *An investigation of the impact of the 6+1 trait writing model on grade 5 student writing achievement* (NCEE 2012-4010). National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Cruz Cordero, T. M. (2024). *A mixed methods exploration of writing motivation among diverse adolescents.* [Doctoral dissertation, University of Delaware]. ProQuest Dissertations & Theses Global. https://udspace.udel.edu/handle/19716/35393

Englehard, G. (1994). Examining rater errors in the assessment of written composition with a many faceted Rasch model. *Journal of Educational Measurement*, 31(2), 93–112. https://doi.org/10.1111/j.1745-3984.1994.tb00436.x

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

Glover, T. A., & Albers, C. A. (2007). Considerations for evaluating universal screening assessments. *Journal of School Psychology*, *45*(2), 117-135. https://doi.org/10.1016/j.jsp.2006.05.005

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29-36. https://doi.org/10.1148/radiology.143.1.7063747

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). John Wiley & Sons, Inc.

Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research*, 60, 237–263. https://doi.org/10.3102/00346543060002237

Jenkins, J. R., Hudson, R. F., & Johnson, E. S. (2007). Screening for at-risk readers in a response to intervention framework. *School Psychology Review, 36*, 582–600.

Juel, C., Griffith, P. L., & Gough, P. B. (1986). Acquisition of literacy: A longitudinal study of children in first and second grade. *Journal of Educational Psychology*, 78(4), 243–255. https://doi.org/10.1037/0022-0663.78.4.243

Keller-Margulis, M., Payan, A., Jaspers, K. E., & Brewton, C. (2016). Validity and diagnostic accuracy of written expression curriculum-based measurement for students with diverse language backgrounds. *Reading and Writing Quarterly*, 32, 174–198. https://doi.org/10.1080/10573569.2014.964352

Kilgus, S. P., Chafouleas, S. M., & Riley-Tillman, T. C. (2013). Development and initial validation of the Social and Academic Behavior Risk Screener for Elementary Grades. *School Psychology Quarterly, 28*(3), 210–226. https://doi.org/10.1037/spq0000024

Kim, Y. S., Al Otaiba, S., Puranik, C., Folsom, J. S., Greulich, L., & Wagner, R. K. (2011). Componential skills of beginning writing: An exploratory study. *Learning and Individual Differences*, *21*(5), 517-525. https://doi.org/10.1016/j.lindif.2011.06.004

Kim, Y. S., Al Otaiba, S., Sidler, J. F., & Gruelich, L. (2013). Language, literacy, attentional behaviors, and instructional quality predictors of written composition for first graders. *Early Childhood Research Quarterly*, *28*(3), 461-469. https://doi.org/10.1016/j.ecresq.2013.01.001

Kim, Y. S. G., & Schatschneider, C. (2017). Expanding the developmental models of writing: A direct and indirect effects model of developmental writing (DIEW). *Journal of Educational Psychology*, *109*(1), 35–50. https://doi.org/10.1037/edu0000129

Kozlow, M., & Bellamy, P. (2004). *Experimental study on the impact of the 6+1 trait writing model on student achievement in writing*. Northwest Regional Educational Laboratory.

McMaster, K. L., Du, X., & Petursdottir, A. (2009). Technical features of curriculum-based measures for beginning writers. *Journal of Learning Disabilities, 42*, 41-60. https://doi.org/10.1177/00222194083262

National Center for Education Statistics. (2012). *What does the NAEP writing assessment measure?* National Center for Educational Statistics, Institute of Education Sciences, U.S. Office of Education. Retrieved from http://nces.ed.gov/nationsreportcard/writing/whatmeasure.aspx

Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing*, 20, 53–76. https://doi.org/10.1016/j.asw.2013.04.001

Shermis, M. D., Burstein, J., Elliot, N., Miel, S., & Foltz, P. (2016). Automated writing evaluation: An expanding body of knowledge. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.). *Handbook of writing research* (pp. 395–409). (2nd ed.). Guilford Press.

Shermis, M. D., & Wilson, J. (2024). Introduction to automated essay evaluation. In M. D. Shermis & J. Wilson (Eds.), *The Routledge international handbook of automated essay evaluation* (pp. 3–22). Routledge.

Shinn, M., Windram, H., & Bollman, K. (2016). Implementing response to intervention in secondary schools. In S. Jimerson, M. Burns, & A. VanDerHeyden (eds) *Handbook of response to intervention*. Springer. https://doi.org/10.1007/978-1-4899-7568-3_32

Silberglitt, B., & Hintze, J. (2005). Formative assessment using CBM-R cut scores to track progress toward success on state-mandated achievement tests: A comparison of methods. *Journal of Psychoeducational Assessment*, *23*(4), 304-325. https://doi.org/10.1177/073428290502300402

Smarter Balanced Assessment Consortium. (2019). *Smarter Balanced Assessment Consortium: 2017–18 summative technical report.* Retrieved from https://portal.smarterbalanced.org/library/en/2017-2018-interim-assessments-technical-report.pdf

Smarter Balanced Assessment Consortium. (2020). *2020-21 Summative Technical Report.* Retrieved from https://technicalreports.smarterbalanced.org/2020-21_summative-report/_book/

Smolkowski, K., & Cummings, K. D. (2015). Evaluation of diagnostic systems: The selection of students at risk of academic difficulties. *Assessment for Effective Intervention*, *41*, 41–54. https://doi.org/10.1177/1534508415590386

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.

Stiggins, R. J. (1982). A Comparison of direct and indirect writing assessment methods. *Research in the Teaching of English*, *16*(2), 101–114. http://www.jstor.org/stable/40170937

Swets, J., & Picket, R. (1982). *Evaluation of diagnostic systems: Methods from signal detection theory*. Academic Press.

Van Norman, E. R., Nelson, P. M., Klingbeil, D. A., Cormier, D. C., & Lekwa, A. J. (2019). Gated screening frameworks for academic concerns: The influence of redundant information on diagnostic accuracy outcomes. *Contemporary School Psychology, 23*, 152–162. https://doi.org/10.1007/s40688-018-0183-0

Whitaker, D., Berninger, V., Johnston, J., & Swanson, H. L. (1994). Intraindividual differences in levels of language in intermediate grade writers: Implications for the translating process. *Learning and Individual Differences*, *6*(1), 107-130. https://doi.org/10.1016/1041-6080(94)90016-7

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

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

Wilson, J., Huang, Y., Palermo, C., Beard, G., & MacArthur, C. A. (2021). Automated feedback and automated scoring in the elementary grades: Usage, attitudes, and associations with writing outcomes in a districtwide implementation of MI Write. *International Journal of Artificial Intelligence in Education, 31*, 234–276. https://doi.org/10.1007/s40593-020-00236-w

Wilson, J., Olinghouse, N. G., McCoach, D. B., Andrada, G. N., & Santangelo, T. (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., Palermo, C., & Wibowo, A. (2024). Elementary English learners' engagement with automated feedback. *Learning and Instruction, 91*, Article 101890. https://doi.org/10.1016/j.learninstruc.2024.101890

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