Assessment in Education: Principles, Policy & Practice

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/caie20

Investigating the promise of automated writing evaluation for supporting formative writing assessment at scale

Joshua Wilson, Matthew C. Myers & Andrew Potter

To cite this article: Joshua Wilson, Matthew C. Myers & Andrew Potter (2022): Investigating the promise of automated writing evaluation for supporting formative writing assessment at scale, Assessment in Education: Principles, Policy & Practice, DOI: 10.1080/0969594X.2022.2025762

To link to this article: https://doi.org/10.1080/0969594X.2022.2025762

ASSESSMENT IN EDUCATION: PRINCIPLES, POLICY & PRACTICE
https://doi.org/10.1080/0969594X.2022.2025762

Routledge
Taylor & Francis Group

Investigating the promise of automated writing evaluation for supporting formative writing assessment at scale

Joshua Wilson, Matthew C. Myers and Andrew Potter

School of Education, University of Delaware, Newark, DE, USA

ABSTRACT

We investigated the promise of a novel approach to formative writing assessment at scale that involved an automated writing evaluation (AWE) system called *MI Write*. Specifically, we investigated elementary teachers' perceptions and implementation of MI Write and changes in students' writing performance in three genres from Fall to Spring associated with this implementation. Teachers in Grades 3–5 (n = 14) reported that MI Write was usable and acceptable, useful, and desirable; however, teachers tended to implement MI Write in a limited manner. Multilevel repeated measures analyses indicated that students in Grades 3–5 (n = 570) tended not to increase their performance from Fall to Spring except for third graders in all genres and fourth graders' narrative writing. Findings illustrate the importance of educators utilising scalable formative assessments to evaluate and adjust core instruction.

ARTICLE HISTORY
Received 31 March 2021
Revised 20 November 2021
Accepted 6 December 2021

KEYWORDS
Writing assessment; elementary; automated essay scoring; automated writing evaluation

Recent attention has focused on identifying effective writing instruction methods at the elementary grades (see Graham et al., 2012). One such method is formative assessment (2015): assessment conducted for the purpose of adjusting and improving instruction and learning (i.e. assessment *for* learning), and not for the purpose of summative evaluation (i.e. assessment *of* learning [see, Bennett, 2011]).

However, teachers face several challenges with conducting formative writing assessment, including generating timely and reliable assessment data (Wilson et al., 2019). These challenges are compounded when formative assessment is conducted at scale, such as when assessing an entire grade level, school, or school district multiple times a year to evaluate students' progress acquiring core knowledge and skills related to academic standards.

Absent a feasible and effective method of formatively assessing writing performance at scale to evaluate students' response to core instruction, educators are left without a means for making timely instructional adjustments to improve core writing instruction or instantiating a 'cycle of continuous improvement' (Bulkley et al., 2010). Year-end summative data are too coarse and come too late.

CONTACT Joshua Wilson joshwils@udel.edu School of Education, University of Delaware, Newark, DE, USA
Supplemental data for this article can be accessed here

This study investigated the promise of a novel approach to scalable formative writing assessment that involved the use of an automated writing evaluation (AWE) system called *MI Write* (www.miwrite.net) that provides immediate and reliable automated scoring coupled with automated feedback.

Automated writing evaluation

There is interest in examining the utility of formative AWE systems to support educational decision making and students' writing development. First, AWE provides efficient and highly reliable scoring without the need for human time and effort. Wilson et al. (2019) reported that assessing students in Grades 3–5 with automated scoring via three 30 min writing prompts, one prompt in each of three genres (narrative, informative, persuasive), yielded reliability coefficients exceeding .80 in each grade level. Moreover, automated scoring retains its scale and reliability over time unlike human scoring (see, Eckes, 2012), affording the ability to describe changes in students' performance across a school year reliably.

Second, AWE feedback has been shown in meta-analysis to improve students' writing quality with an average weighted effect size of 0.38 (Graham et al., 2015). Thus, were AWE to be leveraged for formative assessment at scale it may not only index writing development, but support it (see, Huang & Wilson, 2021; Wilson et al., 2021).

Researchers have begun exploring the promise of automated scoring to support certain key formative assessment decisions. Mercer et al. (2019) found that an automated scoring model based on features extracted from the Coh-Metrix web tool (McNamara et al., 2014) applied to 7-min narrative writing samples had validity coefficients equal to human-scored measures, suggesting the possibility of substituting automated scoring for human scoring for curriculum-based measurement (CBM). These researchers also documented that automated scoring methods applied to 3-min narrative CBM probes had equal or higher diagnostic accuracy than rater-scored methods for identifying students who were non-responsive to core instruction (Keller-Margulis et al., 2021). Similarly, Wilson (2018) and Wilson and Rodrigues (2020) demonstrated that automated scoring applied to 45-min and 30-min writing prompts, respectively, yielded acceptable overall diagnostic accuracy for screening upper-elementary students at risk of writing failure. Collectively, these studies suggest that automated scoring may work well for screening across a wide range of prompt administration times.

Present study

Although research has investigated the potential of AWE to support certain aspects of formative assessment at scale, such as screening, no prior research has explored the potential for AWE to describe changes in students' writing performance across time, a key function of assessments designed to formatively evaluate core instruction (Perie et al., 2009).

Indeed, students exhibit developmental differences in writing skills, (e.g. Abbott & Berninger, 1993; Graham et al., 1997) and different rates of change based on age and initial performance (Keller-Margulis et al., 2015; McMaster et al., 2017): younger and/or lower-performing writers tend to grow faster. Albeit, in absence of effective core instruction, students' growth in writing performance may be stagnant (Keller-Margulis et al., 2016;

Wood et al., 2020). If, via effective formative assessment, educators were to obtain timely, reliable information about changes in students' writing performance, they might be able to intensify core instruction before students' writing difficulties become too severe.

Yet, formatively examining changes in students' writing performance in a single genre is insufficient; academic standards emphasise that students must develop proficiency composing in multiple genres. Different genres impose different constraints on the writer (Donovan & Smolkin, 2006) and students' performance typically differs across genres (Graham et al., 2016). Therefore, students' growth in writing performance in one genre may not be indicative of growth in another genre (McMaster & Campbell, 2008). It is also possible that genre and developmental differences interact, with certain grades exhibiting different rates of change across different genres.

Thus, the present descriptive study involved a year-long naturalistic implementation of MI Write, a formative AWE system. We first described teachers' perceptions and implementation of MI Write, and then described the associated growth in students' writing performance in three genres between Fall and Spring of one school year. Relationships between implementation and rates of change contribute to our understanding of whether AWE might be used to formatively evaluate core writing instruction at scale. Accordingly, the present study investigated the following research questions (RQ):

RQ1: How did teachers perceive the usability and acceptability, usefulness, and desirability of MI Write as a formative assessment tool?

Based on prior research (Wilson et al., 2021), we predicted that teachers would find MI Write to be usable and acceptable, useful, and desirable.

RQ2: How did teachers in Grades 3–5 implement MI Write within their core instruction?

Based on prior research (Wilson et al., 2021), we predicted that implementation would vary by teacher.

RQ3: Associated with teachers' implementation of MI Write, how did students' writing performance in three genres – narrative, informative, and persuasive – change from Fall to Spring?

Based on prior research (e.g. Keller-Margulis et al., 2015; Keller-Margulis et al., 2016; McMaster et al., 2017; Wood et al., 2020) we hypothesised that students would exhibit small amounts of growth across the school year. We also predicted that Grade 3 students would exhibit greater growth on account of being younger (see, McMaster et al., 2017) and that growth would differ across genres because of genre effects on performance (see, Graham et al., 2016). Finally, based on prior research (see Graham et al., 2015), we predicted that students with more extensive AWE usage would demonstrate greater changes in performance across the school year.

Methods

Study context

This study was conducted in 2017–18 with IRB approval waiving documenting informed consent. Data were collected from students in Grades 3–5 in a large Title 1 elementary school in the US Mid-Atlantic region that participated as a part of their pilot of MI Write. The school and its district did not have standardised ELA curriculum. Instead, they

provided standards-aligned educational resources to teachers who utilised whichever resources they wished. MI Write was provided to teachers as one such resource. Within the school, writing instruction was integrated into the 90 min ELA block. The school had 1:1 devices (Chromebooks).

Participants

Our database initially included all students in Grades 3–5 (n = 615) in the school. However, we removed 45 cases who (a) had no Fall or Spring writing data, or (b) who joined the school after the Fall assessment window. The final database consisted of 570 students from 24 classrooms (Grade 3 n = 173; Grade 4 n = 188; Grade 5 n = 209). There were no statistically significant differences with respect to gender, race, or disability status between students retained and trimmed. However, the trimmed sample consisted of a higher proportion of English learners (ELs) [$\chi^2_{(1)} = 9.53$, $p = .002$], which was expected because those students received specialised English language services during testing times.

Demographic information is presented in Table 1. There were no statistically significant differences across grade levels with respect to gender [$\chi^2_{(2)} = 0.54$, $p = .763$], race [$\chi^2_{(6)} = 6.58$, $p = .583$], EL status [$\chi^2_{(2)} = 2.40$, $p = .301$], or disability status [$\chi^2_{(2)} = 0.98$, $p = .611$].

Table 1. Sample Demographics.

	Full Sample	Grade 3	Grade 4	Grade 5
N	570	173	188	209
Gender (%)				
Male	49.1	49.1	51.1	47.4
Female	50.9	50.9	48.9	52.6
Race (%)				
Asian	4.6	4.6	3.2	5.9
Black	50.9	49.7	52.9	50.2
Hispanic	17.7	20.8	13.3	19.1
White	25.3	23.7	28.4	23.3
Other/Two or More Races	1.6	1.2	2.1	1.5
English Learners (%)	9.6	11.6	10.6	7.2
Students with Disabilities (%)	10.7	12.1	11.2	9.1

Measures

Perceptions of MI Write

We utilised an anonymous electronic survey delivered via GoogleForms to gauge teachers' perceptions regarding the adequacy of the amount of time available for instruction and their preparation for teaching writing (3 items), of MI Write's *usability and acceptability* (6 items; α = .84); *usefulness*–measured via two subscales: effects on student learning (7 items; α = .92) and effects on instruction (6 items; α = .90); and *desirability* (1 item; *I would like to continue using MI Write*). All items were rated on a 0 (strongly disagree) to 4 (strongly agree) Likert scale. All items are presented in Appendix A. 14 teachers responded to the survey (five teachers in Grades 3 and 5, and four teachers in Grade 4), a 58% response rate

Implementation of MI Write

We utilised MI Write log data to characterise teachers' and students' implementation and use of MI Write. To describe MI Write implementation at the teacher/classroom level, we calculated the *Total Number of Prompts Assigned* as well as the number of prompts assigned in each genre (narrative, informative, and persuasive). We calculated a second version of each of the four former variables that summarised the number of prompts for which 80% or more of the classroom responded (e.g. *Total Number of Prompts Assigned with 80% Participation, Total Number of Narrative Prompts Assigned with 80% Participation*, etc.). We wished to distinguish use of MI Write to support core instruction from its use to support independent practice. We reasoned that if 80% of students in a classroom submitted an essay in response to a prompt, that prompt could be considered as used for core instruction–we deemed 80% an appropriate criterion because 80% of students are expected to benefit from core curriculum (Fletcher & Vaughn, 2009).

To describe MI Write usage at the student level, we calculated the *Total Number of Prompts Completed*, excluding those used for training and assessment.

Performance

Students were administered randomly assigned writing prompts in three genres in Fall and Spring from a bank of 18 researcher-developed writing prompts (6 prompts in each genre). Appendix B lists all prompts.

Prompt administration procedures. To control for order and topic effects, we counterbalanced genre order across classrooms and randomly assigned prompt topics. Paper copies of prompts were distributed. Then, trained research assistants reminded students how to login to MI Write and locate the appropriate writing activity. Next, they read standardised directions that explained the purpose of the writing activity and key genre features to include when writing. Students were given 30 minutes to write a single draft that they submitted to MI Write at the end of the session.

All prompt administration sessions were audio recorded to ensure fidelity of assessment administration. Twenty percent of the 144 recordings were randomly selected for evaluation. Fidelity of assessment administration, calculated as the percentage of directions read correctly and correct time provided for composing, was high: 93%.

PEG automated scoring. MI Write is a web-based AWE system developed by Measurement Incorporated intended for Grades 3–12. MI Write utilises an automated scoring engine called PEG (Page, 2003). PEG utilises 15 scoring algorithms to evaluate three genres of writing (narrative, informative, and persuasive/argumentative) across five grade bands (Grades 3–4, 5–6, 7–8, 9–10, and 11–12). PEG's scoring algorithms may be reliably applied to any prompt (i.e. prompt-independent scoring algorithms). A detailed description of PEG's architecture is found in Wilson et al. (2021). PEG generates six trait scores to align with the Six Trait model, scoring the traits of *development of ideas, organisation, style, sentence fluency, word choice*, and *conventions* each on a 1.0–5.0 scale. PEG also reports a holistic *Overall Score*, which is the sum of the trait scores (range = 6.0–30.0).

Evaluations of the consistency of PEG with human ratings applied to a test set yielded an average quadratic weighted kappa (QWK) for the three genre models for Grades 3–4 and for the three genre models for Grades 5–6 of 0.87 (SD = 0.03) and 0.85 (SD = 0.04), respectively.

We used the PEG Overall Score for each genre rather than the trait scores. First, conceptually, the measure represents holistic quality, which more closely reflects a student's progress towards achieving grade-level writing standards than any single trait. Second, the Overall Score has a wider scoring range, likely making it more sensitive and useful for describing changes in performance over time. Third, Pearson correlations among PEG's six trait scores were high in Fall and Spring for narrative (.81–.99 and .74–.99), informative (.94 – .99 and .86–.99), and persuasive (.95–.99 and .87–.99) genres. These correlations indicate unidimensionality, suggesting that changes in one trait were likely manifested holistically, rather than uniquely (see, also Keller-Margulis et al., 2021; Wilson, 2018).

Procedures

The first author trained teachers to use MI Write during 45 min team meetings in November 2017. The first author and a team of research assistants trained students to use MI Write during a 45 min class. Then, during three successive days in mid-November students completed the Fall prompts, one prompt in each genre. Following the Fall assessment window, teachers utilised MI Write to support their writing instruction. Finally, over three days in April 2018, we administered the three Spring prompts.

Data analysis

Descriptive statistics summarised teachers' perceptions and implementation of MI Write, our first and second research questions, respectively. To answer our third research question, we conducted a series of multilevel repeated measures analyses wherein repeated measurements of writing quality for each genre at Fall and Spring (i.e. Time; level 1) were nested within students (level 2) who were nested within classrooms (level 3). This multilevel analysis controlled for differences in teacher instruction and MI Write usage. We conducted nine analyses, one for each grade and genre combination.

The multilevel repeated measures model included a dummy coded, uncentered slope parameter which reflected difference in scores from Fall (Time = 0) to Spring (Time = 1; i.e. change scores). At level 2, to predict the Time variable, the MI Write usage variable for PROMPTS was entered using grand-mean centring. Thus, the level-2 intercept represented writing quality scores in the Fall, and the level-2 slope represented the rate of change for students who completed the average number of prompts for their grade. The level-3 model included a classroom level intercept and random error, but no predictors because the student-level usage variable reflected the influence of teacher-level usage. We estimated random effects for the Time slope at the student level (Level 2) but not at the teacher level (Level 3) unless the intra-class correlation (ICC) of the null model indicated that the variance at this level was statistically significantly different from zero. The full mixed model was $Y_{tij} = \gamma_{000} + \gamma_{100}{}^{*}TIME_{tij} + \gamma_{110}{}^{*}TIME_{tij}{}^{*}\text{PROMPTS}_{ij} + r_{0ij} + r_{1ij}{}^{*}TIME_{tij} + u_{00j} + [u_{10j}{}^{*}TIME_{tij}] + e_{tij}$ – the brackets indicate that the random effect for Time at the teacher level was added only under the condition previously described.

Handling of missing data

A total of 439 students (77%) had complete data. Accordingly, we utilised multiple imputation under the assumption that data were missing at random. We utilised SPSS v.26 to create 20 multiply imputed datasets where missing PEG Overall Score data was imputed using a fully conditional Markov chain Monte Carlo method (200 iterations). The imputation model included grade, gender, ELL status, and disability status as predictors, and PEG Overall Scores as both predictors and foci for the imputation.

Software

Multilevel analyses were conducted with Mplus version 8.6 using maximum likelihood estimation with robust standard errors and using the 20 multiply imputed datasets. Pooled results were reported using Rubin's (1987) rules.

Results

Perceptions of MI Write

When asked whether there was sufficient time devoted to writing instruction in their school schedule, teachers were neutral on average but reported a wide range of agreement (M = 2.00; SD = 0.96; range = 0.0–3.0). Teachers also were neutral on average regarding feeling adequately prepared to teach writing, but there was a wide range of responses (M = 2.43; SD = 1.22; range = 0.0–4.0). Teachers agreed that they felt adequately trained to teach writing with MI Write (M = 3.14; SD = 0.66; range = 2.0–4.0).

Based on the aggregate scale, teachers rated the usability and acceptability of MI Write favourably, with an average rating of 2.75 (SD = 0.29; range = 2.29–3.07). However, their perceptions of the acceptability of MI Write's automated scores were more variable, specifically regarding the correlation of MI Write's scores with standards-aligned rubrics (M = 2.57; SD = 1.16) and for predicting scores on the state ELA assessment (M = 2.29; SD = 0.73).

Based on the aggregate subscale, teachers rated MI Write as useful for having positive effects on student learning, with an average rating of 2.84 (SD = 0.21; range = 2.57–3.14) on the seven-item subscale. Teachers also rated MI Write as useful for having positive effects on teachers' instruction, with an average rating of 2.64 (SD = 0.31; range = 2.21–3.00) on the six-item subscale. However, they agreed less strongly that they were effectively using MI Write to teach writing (M = 2.21; SD = 1.05).

Finally, teachers rated MI Write as socially desirable, agreeing that they wished to continue using it in the following year (M = 3.07; SD = 0.83; range = 2.0–4.0).

Implementation of MI Write

Table 2 shows the total number of teacher-assigned prompts during the instructional period. Some teachers used MI Write frequently (e.g. teacher 14) while others used it infrequently or not at all (e.g. teachers 8 and 21). There was also

Table 2. Total Teacher-Assigned Prompts by Genre Across Classrooms During the Instructional Period.

	Teacher	Narrative	Informative	Persuasive	Total
Grade 3	1	1	1	4	6
	2	1	0	4	5
	3	2	3	2	7
	4	0	1	3	4
	5	0	3	4	7
	6	0	0	1	1
	7	0	1	4	5
	8	0	0	0	0
Total		*4*	*9*	*22*	*35*
Grade 4	9	3	0	0	3
	10	2	2	1	5
	11	2	1	1	4
	12	1	1	1	3
	13	2	0	0	2
	14	5	3	1	9
	15	2	5	0	7
	16	3	2	2	7
Total		*20*	*14*	*6*	*40*
Grade 5	17	1	3	1	5
	18	0	2	2	4
	19	0	2	0	2
	20	0	0	4	4
	21	0	0	0	0
	22	0	2	1	3
	23	2	3	2	7
	24	0	3	2	5
Total		*3*	*15*	*12*	*30*
Grand Total		27	38	40	105

variance in usage across genres. For example, teacher 9 assigned only narrative prompts, while teacher 20 only assigned persuasive prompts. Others assigned all three genres nearly equally (e.g. teachers 10 and 23).

There was also variance in genre usage by grade level. Table 3 reports these percentages calculated from frequencies in Table 2. Grade 3 teachers assigned a higher percentage of persuasive prompts, Grade 4 teachers assigned more narrative prompts, and Grade 5 teachers assigned more informative prompts.

Table 3. Percentage of Total Prompts Assigned in Each Genre by Grade During the Instructional Period.

Grade	Narrative	Informative	Persuasive
3	11.4	25.7	62.9
4	50.0	35.0	15.0
5	10.0	50.0	40.0

As indicated in Table 4, usage data suggests that teachers used MI Write to support independent practice versus whole-group instruction. However, Grade 5 teachers utilised proportionately more prompts for whole-group instruction (50%

Table 4. Teacher-Assigned Prompts Completed as Whole-Group Activities.

		Narrative		Informative		Persuasive		Total	
	Teacher	Assigned	80% Completed	Assigned	80% Completed	Assigned	80% Completed	Assigned	80% Completed
Grade 3	1	1	*0*	1	*1*	4	*2*	6	*3*
	2	1	*1*	0	*0*	4	*1*	5	*2*
	3	2	*1*	3	*0*	2	*0*	7	*1*
	4	0	*0*	1	*1*	3	*2*	4	*3*
	5	0	*0*	3	*0*	4	*1*	7	*1*
	6	0	*0*	0	*0*	1	*0*	1	*0*
	7	0	*0*	1	*0*	4	*4*	5	*4*
	8	0	*0*	0	*0*	0	*0*	0	*0*
Total		*4*	*2*	*9*	*2*	*22*	*10*	*35*	*14*
Grade 4	9	3	*0*	0	*0*	0	*0*	3	*0*
	10	2	*1*	2	*0*	1	*0*	5	*1*
	11	2	*1*	1	*0*	1	*0*	4	*1*
	12	1	*1*	1	*0*	1	*0*	3	*1*
	13	2	*0*	0	*0*	0	*0*	2	*0*
	14	5	*1*	3	*2*	1	*0*	9	*3*
	15	2	*1*	5	*0*	0	*0*	7	*1*
	16	3	*2*	2	*1*	2	*1*	7	*4*
Total		*20*	*7*	*14*	*3*	*6*	*1*	*40*	*11*
Grade 5	17	1	*0*	3	*0*	1	*1*	5	*1*
	18	0	*0*	2	*1*	2	*1*	4	*2*
	19	0	*0*	2	*0*	0	*0*	2	*0*
	20	0	*0*	0	*0*	4	*2*	4	*2*
	21	0	*0*	0	*0*	0	*0*	0	*0*
	22	1	*0*	1	*1*	1	*0*	3	*1*
	23	2	*2*	3	*3*	2	*1*	7	*6*
	24	0	*0*	3	*1*	2	*2*	5	*3*
Total		*3*	*2*	*15*	*6*	*12*	*7*	*30*	*15*
Grand Total		27	*11*	38	*11*	40	*18*	105	*40*

of assigned prompts) than teachers in Grades 3 and 4 (40% and 28% of assigned prompts, respectively). Persuasive prompts were used more for whole-group instruction (45% of 40 assigned prompts) than narrative (41% of 27 assigned prompts) and informative prompts (29% of 30 assigned prompts).

Multilevel repeated measures results

Descriptive statistics for the PEG Overall Score across grades, genres, and occasions are presented in Table 5. Results of the multilevel repeated measures analyses are reported separately by grade, below.

Table 5. Descriptive Statistics.

	Grade 3 (n = 173)		Grade 4 (n = 188)		Grade 5 (n = 209)	
	Fall	Spring	Fall	Spring	Fall	Spring
PEG Overall Score Narrative	12.80 (3.40)	13.86 (4.19)	14.75 (3.86)	15.52 (4.27)	15.18 (3.90)	15.06 (3.96)
PEG Overall Score Informative	11.61 (3.90)	13.00 (4.20)	13.94 (4.17)	14.38 (4.54)	14.17 (4.13)	13.98 (4.15)
PEG Overall Score Persuasive	10.47 (3.48)	11.49 (3.92)	12.86 (4.15)	12.61 (4.32)	13.38 (4.15)	13.54 (3.80)

Pooled estimates using Rubin's (1987) rules. Standard deviations are in parentheses.

Grade 3 results

For each genre, intra-class correlations (ICCs) indicated that between 33.18%–40.94% of the variance was within students across time (level 1), 57.60%–65.57% of the variance was between students (level 2), and 0.24%–1.46% of the variance was between classrooms (level 3). In all cases the level-3 variance was not statistically significantly different from 0; therefore, a random effect for Time was not added at this level, only at the student level.

Results of the final, conditional multilevel models for each of the three outcome measures for Grade 3 are presented in Table 6. Students' initial scores in the Fall were highest for narrative (β_{00} = 12.83; SE = 0.22) and lowest for persuasive (β_{00} = 10.47; SE = 0.31). In each genre, third graders exhibited statistically significant positive changes in their writing performance, with the slope term ranging from 0.99 (SE = 0.21) for narrative writing to 1.43 (SE = 0.43) for informative writing. The number of completed prompts was not a statistically significant predictor of students' rate of change in any genre. Finally, there was statistically significant variability in the student-level growth slopes for the narrative genre, but not the other genres.

Grade 4 results

For each genre, ICCs indicated that between 30.13%–39.66% of the variance was at level 1, 50.23%–59.48% of the variance was at level 2, and 7.60%–12.10% of the variance was at level 3. Classroom variance was statistically significant for the informative and persuasive genres; therefore, a random effect for Time was added at this level for those genres.

Table 6. Results of Multilevel Repeated Measures Analyses for Grade 3 Across Narrative, Informative, and Persuasive Genres.

		PEG Overall Score Narrative		PEG Overall Score Informative		PEG Overall Score Persuasive	
Fixed Effect	*df*	Coefficient (*S.E.*)	*t*-ratio	Coefficient (*S.E.*)	*t*-ratio	Coefficient (*S.E.*)	*t*-ratio
Intercept 1, π_0							
Intercept 2, β_{00}	8	12.83 (0.22)	57.52***	11.58 (0.35)	33.49***	10.47 (0.31)	33.95***
Time slope, π_1							
Intercept 2, β_{10}	172	0.99 (0.21)	4.74***	1.43 (0.43)	3.31***	1.01 (0.29)	3.48***
Completed Prompts, γ_{110}	172	0.25 (0.14)	1.82	0.14 (0.15)	0.91	0.12 (0.11)	1.01
Random Effect		Variance Component			Variance Component		Variance Component
Level-1 intercept, *e*		3.20***			5.16***		2.96**
Level 2 intercept, r_0	173	8.30***			9.90***		8.96***
Level 2 slope, r_1	173	5.12***			2.14		2.90
Level 3 intercept, u_{00}	8	0.03			0.03		0.12
Level 3 slope, u_{10}	8						

*$p \leq .05$; **$p \leq .01$; ***$p \leq .001$.

Results of the final, conditional multilevel models for each of the three outcome measures for Grade 4 are presented in Table 7. Students' initial scores in the Fall were highest for narrative (β_{00} = 14.72; *SE* = 0.39) and lowest for persuasive (β_{00} = 12.82; *SE* = 0.50). Fourth graders exhibited statistically significant positive changes in their performance only for narrative writing (β_{10} = 0.78; *SE* = 0.39). The number of completed prompts was not a statistically significant predictor of students' rate of change in narrative or informative writing but did positively predict rate of change for persuasive writing: a one-unit change in the number

Table 7. Results of Multilevel Repeated Measures Analyses for Grade 4 Across Narrative, Informative, and Persuasive Genres.

		PEG Overall Score Narrative		PEG Overall Score Informative		PEG Overall Score Persuasive	
Fixed Effect	*df*	Coefficient (*S.E.*)	*t*-ratio	Coefficient (*S.E.*)	*t*-ratio	Coefficient (*S.E.*)	*t*-ratio
Intercept 1, π_0							
Intercept 2, β_{00}	8	14.72 (0.39)	37.48***	13.90 (0.53)	26.05***	12.82 (0.50)	25.59***
Time slope, π_1							
Intercept 2, β_{10}	187	0.78 (0.39)	2.01*	0.41 (0.51)	0.81	−0.26 (0.28)	−0.92
Completed Prompts, γ_{110}	187	0.26 (0.17)	1.58	0.17 (0.15)	1.18	0.22 (0.11)	2.04*
Random Effect		Variance Component			Variance Component		Variance Component
Level-1 intercept, *e*		4.71***			6.23***		5.06***
Level 2 intercept, r_0	188	9.45***			9.26***		10.26***
Level 2 slope, r_1	188	1.30			0.82		0.46
Level 3 intercept, u_{00}	8	0.88			1.66		1.45
Level 3 slope, u_{10}	8				1.50		0.17

*$p \leq .05$; **$p \leq .01$; ***$p \leq .001$.

of completed prompts was associated with a 0.22 (SE = 0.11) increase in performance in the spring. Variability in student-level and classroom-level growth slopes was not statistically significant.

Grade 5 results

For each genre, ICCs indicated that between 28.35%–36.67% of the variance was at level 1, 56.11%–67.15% of the variance was at level 2, and 4.50%–8.28% of the variance was at level 3 and not statistically significantly different from 0. Therefore, a random effect for Time was not added at this level.

Results of the final, conditional multilevel models for each of the three outcome measures for Grade 5 are presented in Table 8. Students' initial scores in the Fall were highest for narrative (β_{00} = 15.19; SE = 0.40) and lowest for persuasive (β_{00} = 13.45; SE = 0.57). Fifth graders did not exhibit statistically significant positive changes in their writing performance for any genre. The number of completed prompts was not a statistically significant predictor of students' rate of change in narrative or persuasive writing but did predict rate of change for informative writing: a one-unit change in the number of completed prompts was associated with a decrease in informative writing performance in the spring by −0.30 points (SE = 0.08). Variability in student-level growth slopes was not statistically significant.

Discussion

The purpose of the present study was to explore the promise of AWE to support formative writing assessment of core instruction at scale. Teachers generally perceived MI Write to be usable, useful, and desirable, but despite reporting feeling adequately

Table 8. Results of Multilevel Repeated Measures Analyses for Grade 5 Across Narrative, Informative, and Persuasive Genres.

		PEG Overall Score Narrative		PEG Overall Score Informative		PEG Overall Score Persuasive	
Fixed Effect	*df*	Coefficient (*S.E.*)	*t*-ratio	Coefficient (*S.E.*)	*t*-ratio	Coefficient (*S.E.*)	*t*-ratio
Intercept 1, π_0							
Intercept 2, β_{00}	8	15.19 (0.40)	37.65***	14.18 (0.58)	24.50***	13.45 (0.57)	23.51***
Time slope, π_1							
Intercept 2, β_{10}	208	−0.13 (0.34)	−0.37	−0.20 (0.46)	−0.43	0.01 (0.31)	0.03
Completed Prompts, γ_{110}	208	0.14 (0.17)	0.81	−0.30 (0.08)	−3.68***	−0.06 (0.07)	−0.90
Random Effect		Variance Component			Variance Component		Variance Component
Level-1 intercept, e		4.10***			5.74***		4.81***
Level 2 intercept, r_0	209	10.35***			9.51***		9.53***
Level 2 slope, r_1	209	0.46			1.12		0.28
Level 3 intercept, u_{00}	8	0.68			1.31*		1.31*
Level 3 slope, u_{10}	8						

*$p \leq .05$; **$p \leq .01$; ***$p \leq .001$.

trained, teachers agreed less strongly that they were effectively using MI Write to teach writing. Moreover, implementation data indicate potential contradictions between teachers' perceptions and utilisation of the software.

For example, teachers perceived MI Write to be easy to use and to increase students' opportunities to receive feedback and revise, yet about half created just one prompt per month during the usage period. Further, it appeared that teachers elected not to implement MI Write primarily to support core instruction: of the 105 total prompts assigned, only 40 were completed by 80% or more of students. Reasons for this disjuncture between teachers' perceptions and behaviour are unclear but may have been due to availability of sufficient time devoted to writing instruction in their schedule or their preparation for teaching writing, as suggested by teachers' survey responses. Thus, in the context of a school that did not have a standardised ELA curriculum, implementation of MI Write was rather limited.

Associated with this level of implementation, results of multilevel repeated measures analyses indicated that students' writing performance in three genres tended not to change between Fall and Spring, albeit with two exceptions: third graders increased their writing performance in all three genres and fourth graders increased their narrative performance. Moreover, there was statistically significant variability in the growth slope only for Grade 3 narrative writing.

These results are consistent with prior research indicating that in absence of effective core instruction students exhibit little or no change in writing performance over time, particularly in later elementary grades (Wood et al., 2020), and little variability in within-year growth (Keller-Margulis et al., 2016). In the present study, teachers had no standard writing curriculum and there was no dedicated writing block. Also, teachers' use of MI Write reflected a rather weak implementation. Absent sufficiently rigorous core instruction it is reasonable that students' writing performance tended not to improve. Findings underscore the importance of access to scalable and effective formative writing assessments to formatively evaluate, and subsequently adapt, core instruction.

Grade-level differences in rates of change may be partially explained by teachers' MI Write implementation. Third graders displayed statistically significant positive changes in writing performance and third-grade teachers assigned the second highest number of prompts within MI Write (35 total, an average of 4.375 prompts per teacher) and second highest number of prompts completed with 80% class participation (14 out of 35; 40%). However, fourth-grade teachers assigned more prompts in total (40, an average of 5 prompts per teacher) and fifth-grade teachers had the highest the number of prompts completed with 80% class participation (15 out of 30; 50%), yet fourth graders only increased their narrative writing performance, and fifth graders did not increase their performance in any genre. The increase in fourth grade narrative writing might be explained by the fact that fourth graders primarily used MI Write to compose narrative text (50% of assigned prompts). Nevertheless, usage cannot fully explain study findings.

Alternatively, developmental differences may help explain third graders' positive changes in performance and variability in growth in narrative writing. Prior research has shown that third graders experience greater growth in writing than older students (McMaster et al., 2017; Wood et al., 2020). Further, McMaster et al. (2017) found that initial performance was negatively associated with slope. In the present study third graders had the lowest fall scores for all genres compared to other grades, potentially providing more room to improve.

Interestingly, after accounting for the clustered nature of the data, the amount of writing practice students experienced within MI Write (i.e. the number of completed prompts) was not associated with students' rate of change except for two instances. First, there was a positive association with changes in fourth graders' persuasive writing performance. Fourth graders completed the least amount of persuasive writing prompts (15% of assigned prompts), suggesting that genre-specific practice cannot explain our findings. Fourth-grade teachers assigned the greatest amount of narrative writing prompts, the greatest number of prompts overall (40 prompts assigned), but the least number of prompts assigned as part of core instruction–only 11 of 40 prompts had ≥ 80% class participation. However, fourth graders' writing performance in the Fall was weaker in persuasive writing than the other genres. Thus, the positive association between prompts completed and changes in persuasive writing performance might be explained by an overall practice effect that helped students who completed additional independent practice to improve their weakest area of performance: persuasive writing.

Second, we found a negative association between the number of completed prompts with changes in fifth graders' informative writing performance. Although fifth graders completed the greatest amount of practice in informative writing (50% of assigned prompts), they often experienced this practice outside the context of core instruction: only 6 of 15 prompts had ≥ 80% class participation. Thus, it may be that fifth graders who completed more prompts did not independently identify ways of strengthening their informative writing performance, signalling the need for more explicit, core instruction in informative writing at this grade level (e.g. Hebert et al., 2021).

Limitations and future research

Future research should be conducted in schools with more rigorous instructional contexts. If we didnot see greater increases in students' performance in such a context, we might question the suitability of MI Write for formatively evaluating the effectiveness of core writing instruction.

Second, absent human ratings of the Fall and Spring prompts or a separate standardised measure, it is not possible to corroborate study findings based on MI Write's automated scores. Despite MI Write's consistency with human ratings, students may have demonstrated changes that were not detected by MI Write's automated scoring.

Third, we examined changes in performance using only two points of measurement. Future AWE research should apply growth models to three or more data points.

Fourth, we received survey responses from 14 of the 24 (58.3%) teachers who participated in the study, introducing the possibility of selection bias.

Finally, students may have experienced practice opportunities outside of MI Write that we could not account for in our analyses.

Conclusion

Study findings underscore the importance of educators conducting formative writing assessment at scale. Without generating timely, reliable information about students' changes in performance across time, educators are left without a means of adjusting core instruction if students' exhibit little to no change in writing performance, as we saw

in several cases in the present study. AWE systems, such as MI Write, have several desirable characteristics to support scalable formative assessment: scores are delivered immediately and without the need for human labour, they are 100% consistent, and they retain the same reliability and scale over time. Although more research is needed, study findings are an important first step in expanding the range of formative assessments available to educators to evaluate students' response to core instruction.

Disclosure statement

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

Funding

This research was supported by Grant #201800044 from the Spencer Foundation to the University of Delaware. The opinions expressed are those of the authors and do not represent the views of the Foundation, and no official endorsement by this agency should be inferred. The authors declare no conflicts of interest relative to this research study.

Notes on contributors

Joshua Wilson, Ph.D., is an Associate Professor at the University of Delaware School of Education. His research focuses on ways that automated writing evaluation can help to transform the teaching and learning of writing. A former special education teacher, Dr. Wilson is particularly concerned with improving writing outcomes for those most at risk of learning difficulties.

Matthew C. Myers is a Ph.D. student in Educational Statistics and Research Methods at the University of Delaware School of Education. His research interests include applications of machine learning for writing assessment as well as the evaluation and optimization of machine learning approaches.

Andrew Potter is a Ph.D. student at the University of Delaware School of Education. His research interests include integrated reading and writing instruction and development for at-risk students using quantitative and mixed methods.

ORCID

Joshua Wilson http://orcid.org/0000-0002-7192-3510

References

Abbott, R. D., & Berninger, V. W. (1993). Structural equation modeling of relationships among developmental skills and writing skills in primary- and intermediate-grade writers. *Journal of Educational Psychology*, *85*(3), 478–508. https://doi.org/10.1037/0022-0663.85.3.478

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy, & Practice*, *18*(1), 5–25 . https://doi.org/10.1080/0969594X.2010.513678

Bulkley, K. E., Olah, L. N., & Blanc, S. (2010). Introduction to the special issue on benchmarks for success? Interim assessments as a strategy for educational improvement. *Peabody Journal of Education*, *85*(2), 115–124. https://doi.org/10.1080/01619561003673920

Donovan, C. A., & Smolkin, L. B. (2006). Children's understanding of genre and writing development. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 131–143). The Guilford Press.

Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly*, *9*(3), 270–292. https://doi.org/10.1080/15434303.2011.649381

Fletcher, J. M., & Vaughn, S. (2009). Response to intervention: Preventing and remediating academic difficulties. *Child Development Perspectives*, *3*(1), 30–37. https://doi.org/10.1111/j.1750-8606.2008.00072.x

Graham, S., Berninger, V. W., Abbott, R. D., Abbott, S. P., & Whitaker, D. (1997). Role of mechanics in composing of elementary school students: A new methodological approach. *Journal of Educational Psychology*, *89*(1), 170–182. https://doi.org/10.1037/0022-0663.89.1.170

Graham, S., Hebert, M., & Harris, K. R. (2015). Formative assessment and writing: A meta-analysis. *Elementary School Journal*, *115*(4), 523–547. doi:https://doi.org/10.1086/681947.

Graham, S., Hebert, M., Sandbank, M. P., & Harris, K. R. (2016). Assessing the writing achievement of young struggling writers: Application of generalizability theory. *Learning Disability Quarterly*, *39*(2), 72–82. https://doi.org/10.1177/0731948714555019

Graham, S., McKeown, D., Kiuhara, S., & Harris, K. R. (2012). A meta-analysis of writing instruction for students in the elementary grades. *Journal of Educational Psychology*, *104*(4), 879–896. https://doi.org/10.1037/a0029185

Hebert, M., Bazis, P., Bohaty, J. J., Roehling, J. V., & Nelson, J. R. (2021). Examining the impacts of the structures writing intervention for teaching fourth-grade students to write informational text. *Reading and Writing*, *34*(7), 1711–1740. https://doi.org/10.1007/s11145-021-10125-w

Huang, Y., & Wilson, J. (2021). Using automated feedback to develop writing proficiency. *Computers and Composition*, *62*, 102675. https://doi.org/10.1016/j.compcom.2021.102675

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 & Writing*, *34*(10), 2461–2480. https://doi.org/10.1007/s11145-021-10153-6

Keller-Margulis, M. A., Mercer, S. H., Payan, A., & McGee, W. (2015). Measuring annual growth using written expression curriculum-based measurement: An examination of season and gender differences. *School Psychology Quarterly*, *30*(2), 276–288. https://doi.org/10.1037/spq0000086

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

McMaster, K. L., & Campbell, H. (2008). New and existing curriculum-based writing measures: Technical features within and across grades. *School Psychology Review*, *37*(4), 550–566. https://doi.org/10.1080/02796015.2008.12087867

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-Metrix*. Cambridge University Press.

Mercer, S. H., Keller-Margulis, M. A., Faith, E. L., Reid, E. K., & Ochs, S. (2019). The potential for automated text evaluation to improve the technical adequacy of written expression curriculum-based measurement. *Learning Disability Quarterly*, *42*(2), 117–128. https://doi.org/10.1177/0731948718803296

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.

Perie, M., Marion, S., & Gong, B. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues and Practice*, *28*(3), 1–15. https://doi.org/10.1111/j.1745-3992.2009.00149.x

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons.
Wilson, J. (2018). Universal screening with automated essay scoring: Evaluating classification accuracy in Grades 3 and 4. *Journal of School Psychology*, *68*(1), 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*(4), 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*(2), 234–276. https://doi.org/10.1007/s40593-020-00236-w
Wilson, J., & Rodrigues, J. (2020). Classification accuracy and efficiency of writing screening using automated essay scoring. *Journal of School Psychology*, *82*(1), 123–140. https://doi.org/10.1016/j.jsp.2020.08.008
Wood, C. L., Schatschneider, C., & Hart, S. (2020). Average one year change in lexical measures of written narratives for school age students. *Reading & Writing Quarterly*, *36*(3), 260–277. https://doi.org/10.1080/10573569.2019.1635544