Integrating Goal-Setting and Automated Feedback to Improve Writing Outcomes: A Pilot Study

Joshua Wilson (ORCID#: 0000-0002-7192-3510), Andrew Potter, Tania Cruz Cordero, and Matthew C. Myers

University of Delaware

Author Note

Joshua Wilson, Ph.D., School of Education, University of Delaware; Andrew Potter, M.S., School of Education, University of Delaware; Tania Cruz Cordero, M.S., School of Education, University of Delaware; Matthew C. Myers, M.Ed., School of Education, University of Delaware;

Acknowledgments

The authors declare no conflicts of interest relative to this research study.

Correspondence for this article should be addressed to Joshua Wilson, Ph.D., University of Delaware, School of Education, 213E Willard Hall Education Building, Newark, DE, 19716, United States. Tel: +13028312955. Email: joshwils@udel.edu.

Citation:
Wilson, J., Potter, A., Cruz Cordero, T., & Myers, M. (2022). Integrating goal-setting and automated feedback to improve writing outcomes. *Innovation in Language Learning and Teaching*, DOI: 10.1080/17501229.2022.2077348

Abstract

Purpose: This study presents results from a pilot intervention that integrated self-regulation through reflection and goal setting with automated writing evaluation (AWE) technology to improve students' writing outcomes.

Methods: We employed a single-group pretest-posttest design. All students in Grades 5–8 (*N* = 56) from one urban, all female, public-charter middle school completed pretest and posttest measures of writing beliefs and writing performance. In between pretest and posttest, students completed monthly goal-setting activities via a Qualtrics survey and monthly persuasive writing practice via prompts completed within an AWE system.

Findings: Students improved their self-regulation as indicated by improved goal calibration and confidence to achieve their goals over time. They also improved their self-efficacy for writing self-regulation and writing performance between pre and posttest. Students also perceived the intervention to be usable, useful, and desirable.

Originality: This is a unique study because we integrated AWE and goal-setting instruction, which has not previously been done. Positive findings indicate the promise of this innovative, feasible, and scalable technology-based writing intervention.

Keywords: Automated writing evaluation, automated feedback, writing technology, goal setting, self-regulation, educational technology implementation.

Integrating Goal-Setting and Automated Feedback to Improve Writing Outcomes: A Pilot Study

Writing is a critical skill required for academic success, yet U.S. students perform below proficiency levels on large-scale writing assessments; proficiency gaps are even larger for students from urban locales, minority backgrounds, and those experiencing poverty (National Center for Education Statistics, 2012). Thus, there is a need for innovative writing interventions, especially for those most at risk.

This study presents results from a pilot intervention aimed to improve self-regulation for writing, writing self-efficacy, attitudes toward writing, and writing performance of urban middle school students of color by integrating goal setting with automated writing evaluation (AWE). AWE are software platforms that provide immediate computer-generated feedback via evaluation scores and feedback comments (Strobl et al., 2019). Although both goal setting and AWE have been shown to be effective separate of each other, no prior study has explicitly integrated these two approaches to improve outcomes of interest. Thus, our study explores the promise of an innovative approach to addressing the persistent problem of poor writing performance, especially among those most at risk.

Self-Regulation in Writing

Writing requires self-regulation of cognition (planning, translating, reviewing, and revising) and affect (beliefs and motivation; see Hayes, 1996). According to Zimmerman and Risemberg (1997), writers enact self-regulatory behaviors that are personal (e.g., applying a planning strategy), behavioral (e.g., setting and monitoring progress toward goals), or environmental (e.g., manipulating an environment to reduce distractions). As writers enact these self-regulatory behaviors, they evaluate the effectiveness of those behaviors, continuing or modifying them based on evidence of their effectiveness. In turn, this self-regulatory feedback loop influences a writer's *self-efficacy*—perceptions of one's ability to successfully execute a writing task—and motivation for, and enjoyment of, writing (Zimmerman & Risemberg, 1997).

For example, a student who struggles to self-regulate while writing may have low self-efficacy and low expectations for success. Consequently, this student may hold negative attitudes toward writing. Alternatively, a student who effectively self-regulates when writing may have high self-efficacy gained through mastery experiences (i.e., personal experiences of successfully completing a task and receiving positive feedback [Pajares et al., 2007]). This student is more likely to hold positive expectations about their future success and maintain more positive attitudes toward writing. As these examples illustrate, self-regulation is intertwined with self-efficacy, beliefs, and attitudes.

Developing Self-Regulation Through Goal Setting

One effective method for building self-regulation is *goal setting*—goal setting involves a student making a goal that is specific and appropriately challenging (Reid et al., 2013). Goal setting assists students in developing self-regulation by helping them understand what they need to work toward and by providing motivation/reinforcement as they work toward their established goal (Reid et al., 2013). Results from multiple meta-analyses suggest that goal setting is an evidence-based self-regulatory strategy with a large effect size (ES = 0.80; Graham & Harris, 2018). Moreover, a review of experimental writing self-efficacy studies suggested that a crucial aspect of writing self-efficacy involves setting and monitoring progress toward achieving writing goals (Bruning & Kauffman, 2016).

Goal setting interventions for writing may include setting *process goals* and *product goals* (Ferretti & Fan, 2016; Graham et al., 1992; Torrance et al., 2015). Process goals target how a writer engages in the writing process, including planning, drafting, revising, and editing (e.g., "I will develop a full outline before I draft"). Product goals target improvements in the writing content and quality (e.g., "I will make sure each reason is supported by effective evidence"). Students who set both types of goals improve their writing process and performance (Torrance et al., 2015).

AWE and Goal Setting

Feedback helps students make correct judgements about their performance and develop positive self-efficacy beliefs (see Schunk & Pajares, 2002). In addition to teacher feedback, AWE feedback also promotes these outcomes. Although research on AWE has largely focused on college students learning English as a foreign or second language (Stevenson & Phakiti, 2014), findings from research on AWE with adolescents in the US indicate that AWE is associated with increased writing self-efficacy (Wilson & Roscoe, 2020), writing motivation (Grimes & Warschauer, 2010; Wilson & Czik, 2016), effective revising (Huang & Wilson, 2021; Roscoe et al., 2018), use of text evidence (Wang et al., 2020), and improved writing performance (Graham et al., 2015; Palermo & Thomson, 2018). In addition, teachers and students have reported positive perceptions about AWE's *social validity* (Roscoe et al., 2014; Wang et al., 2020; Wilson & Roscoe, 2020)—social validity refers to stakeholder perceptions of the usability, usefulness, and desirability of an intervention (Lyst et al., 2005; Wolf, 1978). Interventions with greater social validity are more likely to be implemented and sustained.

Nevertheless, with the exception of the Writing Pal AWE system that integrates AWE with strategy instruction (Roscoe & McNamara, 2013), one limitation of existing research is that AWE is rarely integrated with other effective writing-instruction practices. One notable exception is a study by Palermo and Thomson (2018), who found that writing instruction was more effective when teachers integrated AWE with self-regulated strategy development instruction compared to AWE in the context of traditional instruction or a control group that did not utilize AWE. Nevertheless, the most common implementation of AWE in the research literature is as a stand-alone tool (c.f., Link et al., 2020; Wilson & Czik, 2016), and no prior study has explored the integration of AWE with goal setting as we did in the current study.

AWE has several affordances that may support goal setting. First, AWE provides immediate feedback in the form of essay ratings and suggestions for improvement. Students can use that feedback to form mastery experiences and thereby improve self-efficacy, positive expectations for future success, and self-regulation. Second, AWE feedback may help students more accurately assess their performance and develop realistic expectations for their writing goals. That is, AWE may help students become better *calibrated*—calibration is "the ability to accurately judge one's performance" (Rutherford, 2017, p.33), an important aspect of self-regulation associated with academic performance (Rutherford, 2017). Third, AWE helps with reliably monitoring progress toward product and process goal via its immediate essay ratings, electronic portfolios, and tools to support the writing process.

Present Study

Our pilot intervention aimed to help students reflect on and set writing product and writing process goals while leveraging AWE's affordances to support goal setting. Drawing on from theoretical work that underscores the importance of self-regulation and from empirical research that shows the positive effects of goal setting and AWE on writing outcomes, this is the first study to directly integrate goal setting and AWE.

We anticipated that, over time, students would gain facility with the process of goal setting and concomitantly develop better self-regulation as evidenced through improved goal calibration, increased confidence to achieve their goals, increased writing self-efficacy, and improved attitudes toward writing. Moreover, given consistent positive relationships between writing self-efficacy and writing performance (Pajares, 2003), as well as positive relationships between writing attitudes and performance among adolescents (see Graham et al., 2018), we hypothesized that students would improve their writing performance. We also hypothesized that students would hold positive perceptions of the social validity of the pilot intervention.

Thus, our study was designed to answer the following research questions (RQs):

1. To what extent did students achieve their product goals? Did students become better calibrated over time?
2. Did students' confidence toward achieving their monthly goals improve over time?
3. Did students improve their writing self-efficacy and attitudes toward writing between pretest and posttest?
4. Did students improve their writing performance between pretest and posttest?
5. To what extend did students perceive the integrated goal setting and AWE feedback pilot intervention to be usable, useful, and desirable?

Method

Research Design

We employed a single group pretest-posttest research design to simultaneously describe changes in students' goal setting behavior and performance over time and to examine the degree to which implementation of our pilot intervention was associated with gains in outcomes of interest.

Participants

All enrolled students (*N* = 56) in Grades 5–8 from one urban, all female, charter middle school in the Mid-Atlantic United States participated in the study (Grade 5 = 16.1% of sample; Grade 6 = 32.1%, Grade 7 = 25.0%; Grade 8 = 26.8%). Students were primarily Black (75.0%) and Hispanic/Latinx (23.2%), and from low-income families (92.9% received free/reduced price lunch). No students received special education services or were English learners. All students were taught by the same English language arts (ELA) teacher. Per IRB approval, parental consent and student assent were obtained via an opt-out procedure; no parents or students opted out of the research study.

Pedagogical Materials

MI Write

MI Write uses the Project Essay Grade (PEG; Page, 2003) scoring engine to provide immediate automated scores and feedback within the Six Traits of Writing framework (i.e., Development of Ideas, Organization, Style, Sentence Fluency, Word Choice, and Conventions; see Coe et al., 2011). MI Write's Overall Score, which is intended to serve as an indicator of holistic writing quality, is formed as the sum of the individual six trait scores (range = 6–30). Further, MI Write provides feedback on grammar and spelling separate from the feedback it provides on the six traits. The platform includes electronic graphic organizers to support planning, interactive lessons, and a peer review function. An electronic portfolio provides opportunities for students to monitor progress and for teachers to support them accordingly. See Appendices A and B for screenshots of MI Write feedback score reports and graphic organizers.

Goal Setting Survey

We created a goal setting survey (GSS) of our own design for students, programmed on Qualtrics, based on prior research on goal setting in writing (e.g., Ferretti & Fan, 2016; Graham et al., 1992; Torrance et al., 2015). The survey had three sections that focused on setting (1) process goals, (2) product goals, and (3) an overall performance goal.

The first section informed students how writers set process goals to address planning, drafting, revising, and/or editing. To scaffold students' reflection, students selected from a list of actions they enacted for each process in their last writing assignment. For instance, for planning, students were asked to consider whether they did the following: "select a planning template from MI Write," "write down key words and ideas for each part of the essay," "elaborate on my ideas briefly," "set my ideas in the order they will follow in the final essay," "double-check my plan before starting my draft," or "none." Similar reflection and selection opportunities were provided for the processes of drafting, revising, and editing. Students could choose all, some, or no actions.

After reflecting, students were asked to choose two writing processes to improve on. For each of the two processes they selected, students were once again shown the list of good practices and were asked "choose a strategy that you didn't use in your last essay. This will be your goal for the next one." We scaffolded goal setting in this way because students with limited genre knowledge and self-regulation skills may benefit from the provision of explicit goals to improve their writing (Ferretti & Fan, 2016).

The second part of the GSS informed students about product goals and helped them reflect on and set product goals in a similar scaffolded process. Students reflected on the MI Write scores they received on the most recent essay they wrote—when completing the GSS for the very first time, students were directed to review their performance on their pretest essay completed in the prior month. Students were directed to enter the scores for each trait in text entry boxes. On a subsequent page, students' scores for each trait were re-displayed and students were prompted to select two traits to improve, such as their two lowest-scoring traits.

Students then viewed a list of four strategies specific to each trait they selected for improvement. In all cases, strategy options were selected from the six-trait feedback provided by MI Write for essays receiving between a 4.0 and 5.0 on that given trait (i.e., scores indicating high performance) but rephrased in "I"-language to match the language of a goal (e.g., "Make sure I state my opinion clearly"). Thus, we built explicit connections between students' product goals and MI Write's feedback for the purpose of helping students understand the evaluation criteria against which their writing would be judged, as helping students understand evaluation criteria is a fundamental principle of formative assessment (Black & Wiliam, 2009).

The final part of the survey directed students to input their prior final-draft Overall Score and then set a goal for their Overall Score for their next essay. To help students set a reasonable goal, we provided with the following statement based on prior research conducted by Palermo and Wilson (2020): "We know that typical improvement between essays is about 3 points, but you can pick any goal you wish." The survey concluded by asking students to rate their confidence to achieve their Overall Score goal between 0 (not at all confident) to 100 (completely confident) using a sliding scale. Students did not provide a confidence rating in the first month (i.e., November).

The final screen of the survey presented the student with a summary of their goals and encouraged them to record those goals for easy access during their next writing assignment.

The GSS is viewable at: https://delaware.ca1.qualtrics.com/jfe/form/SV_9R13tVZPxS3ra2G

Persuasive Essay Prompts

Pretest and Posttests. Students were assigned one of two persuasive prompts at pretest and posttest. The persuasive genre was selected given its prominence in contemporary US state standards and low performance among US students in this genre (Ferretti & Graham, 2019). Prompts were selected by researchers with experience teaching middle school ELA to ensure that students would have adequate background knowledge to respond to the prompts. Cultural responsiveness of the prompt topics was also considered in relation to the study sample. One prompt asked students to write a persuasive argument for or against banning passenger cars in some areas and requiring people to walk, bike, or use public transportation. The other prompt asked students to write a persuasive argument for or against schools replacing meat with vegetarian meals on certain days of the week. Prompt topics were counterbalanced within-subjects across pretest and posttest to control for prompt effects.

Monthly Writing Assignments. Students were assigned one persuasive essay prompt per month between November and March within MI Write (5 total). Prompts were created by the ELA teacher with assistance from the first author and were related to topics of interest within the curriculum and society (e.g., nonviolence and bravery, voices of young women of color). These monthly persuasive essay prompts included sources (e.g., videos or text).

Measures

Writing Beliefs Survey

At pretest and posttest students completed an electronic survey that assessed students' self-efficacy for writing and their attitudes toward writing.

Self-efficacy. The *Self-Efficacy for Writing Scale* (Bruning et al., 2013) prompted students to rate themselves on 19 items from 0 to 100 on how confident they were to engage in different behaviors while writing (Cronbach's $\alpha = .95$ at both pretest and posttest). This measure includes three sub-scales that showed good reliability: *Conventions* (five items; e.g., "I can write complete sentences"; pretest α = .86; posttest α = .89)*, Idea generation* (six items; e.g., "I can think of many ideas for my writing"; pretest α = .88; posttest α = .92), and *Self-regulation* (eight items; e.g., "I can make a good plan for my writing"; pretest and posttest α = .91).

To further evaluate students' perceptions of self-efficacy, we administered an adapted version of the *Ability and Expectancy Beliefs* scale developed by Wigfield and Eccles (2000). This scale requires students to rate their writing abilities relative to their peers (e.g., "If you were to list all the students in your class from the worst to the best in writing, where would you put yourself?"), and their performance expectations in writing (e.g., "How good would you be at learning something good in writing?") on a 0 to 4 scale, with higher values indicating greater confidence in one's ability and expectations for performance. The scale had acceptable reliability at pretest (α = .72) and good reliability at posttest (α = .82).

Attitudes Toward Writing. We measured students' attitudes toward writing (i.e., liking writing) using MacArthur et al.'s (2016) *Affect about Writing* scale. This scale asked students to report their level of agreement with five statements on a Likert scale of 0 (strongly disagree) to 4 (strongly agree) (e.g., "I usually enjoy writing," "The process of writing is satisfying for me"). Reliability was high at pretest (α = .88) and posttest (α = .90).

Writing Performance

Pretest and posttest writing performance was assessed via MI Write's Overall Score of writing quality. Prior research has demonstrated this measure's evidence of score reliability (Shermis, 2014; Wilson et al., 2019) and convergent validity with human scores (Wilson & Czik, 2016). In all analyses we utilized the Overall Score because it represented the construct of overall writing quality and because of multicollinearity between trait scores.

Social Validity Survey

We designed a survey to investigate students' perceptions of the social validity of the pilot intervention. Specifically, at posttest, students rated MI Write and the GSS in terms of usability, usefulness, and desirability. All items ranged between 0 (strongly disagree) and 3 (strongly agree).

MI Write. The usability subscale consisted of five items that asked students how easy it was to use MI Write overall and specific features (e.g., "find the correct writing assignment," "use MI Write's graphic organizers"). The usefulness subscale consisted of six items that asked students to report if MI Write helped them in various ways (e.g., planning, revising, monitoring progress, becoming a better writer). The desirability subscale consisted of two items: "I would recommend MI Write to other students," and "I would like to continue using MI Write." Scale reliability was acceptable for usability ($\alpha = .71$), usefulness ($\alpha = .75$) and desirability ($\alpha = .85$).

Goal Setting Survey. We asked students about the usability of the GSS with one Likert scale item: "Overall it was easy to use the survey." We asked students about the usefulness of the GSS with eight Likert scale items (e.g., "The survey helped me set clear goals for my writing"). Students rated the desirability of the GSS via two items: "I would recommend this goal setting survey to other students," and "I would like to continue using this goal setting survey." Scale reliability was acceptable for usefulness ($\alpha = .79$) and desirability ($\alpha = .84$).

Procedures

This study was conducted from October 2020 to May 2021. In October, students completed a two-session pretest. In session 1, they completed the writing beliefs survey and had 30 minutes to write a first draft of a randomly assigned persuasive prompt in MI Write. A day later during session 2, they revised using MI Write's feedback and submitted the final draft.

Once per month between November 2020 and March 2021, students completed the GSS and subsequently composed a persuasive essay within MI Write. In November, when students completed the GSS for the first time, they watched a training video recorded by the first author and provided in the email they received. Students accessed the GSS via a hyperlink emailed directly to them by the research team. Students logged into MI Write with their individual usernames and passwords. Students completed the GSS and writing assignments independently. The ELA teacher's role was exclusively to assist with technical and procedural issues (e.g., accessing the GSS, logging in to MI Write, moving from prewriting to drafting) and manage the classroom.

In April 2021, students completed a two-session posttest following the same procedures as at pretest but with the addition of the social validity survey in session two.

Trained researchers administered the pretest and posttest remotely over Zoom following a script. Sessions were audio-recorded to confirm fidelity of assessment administration, which was calculated as the percentage of standardized directions that were correctly delivered and correct timings maintained. Fidelity across all administration sessions was high at pretest (M = 97%; SD = 1.7%) and posttest (M = 97%; SD = 2.3%).

Class modality changed because of state mandates during the COVID-19 pandemic. The school held in-person learning from September to December and virtual learning from December to early March. Throughout these changes, some students elected to remain remote. The changes in class modality did not affect the delivery or nature of the intervention because it was designed to occur fully online. However, these changes did contribute to rates of missing data.

Data Analysis

To answer RQ1 and RQ2, we used descriptive statistics and longitudinal multilevel regression models to determine if students improved their calibration accuracy and confidence over time. To answer RQ3 and RQ4, we used descriptive and inferential statistics to examine mean differences between pre and posttest for writing beliefs and performance. We employed paired-sample *t*-tests with a Bonferroni correction to examine differences in writing beliefs and writing performance (measured as MI Write Overall Score). In all cases, we used listwise deletion to handle missing data at posttest. Finally, we used descriptive statistics to answer RQ5 about student perceptions of the intervention.

We experienced missing data on monthly student essays and surveys (see Table 1). We decided to drop data from December in our analyses due to the high percentage of missing data on both measures stemming from school closures and the shortened month due to the holiday break.

Table 1

Percentage of Complete Data by Activity by Month

Month	Percentage of Completed GSS	Percentage of Completed Essays
November	89.29%	62.50%
December	33.93%	33.93%
January	89.29%	76.79%
February	66.07%	48.21%
March	78.57%	73.21%

Note. Number of total participants = 56. GSS = goal setting surveys.

Results

RQ1: Goal Achievement and Calibration

Goal Achievement

Figure 1 displays the percentage of students with available data (see Table 1) that met and did not meet the overall score goal they set for themselves each month. Descriptive statistics for

Accepted Manuscript
Version of record at: https://doi.org/10.1080/17501229.2022.2077348

monthly score goals and actual performance are presented in Table 2. The descriptive statistics presented in Figure 1 and Table 2 indicate that increasing percentages of students met their goals after the first prompt in November, but most students still did not meet their goals by March. However, the mean difference between students' goals and actual achievement generally decreased over time (see Table 2).

Table 2

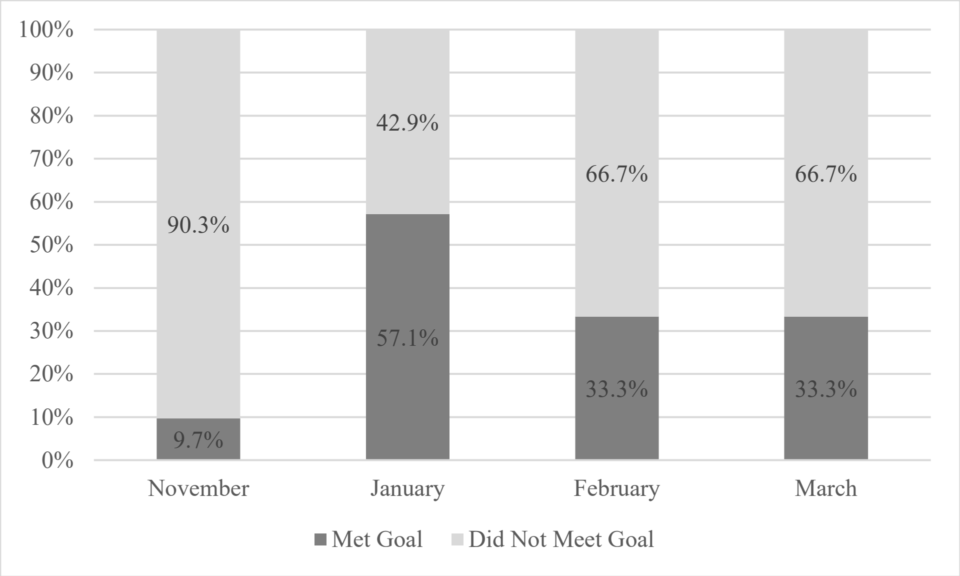
Descriptive Statistics of Monthly Total Scores and Monthly Total Score Goals

Month	Overall Score Goal *M* (*SD*)	Overall Score Achieved *M* (*SD*)	Mean Difference between Goal and Achievement
November	21.32 (6.97)	12.90 (4.13)	8.42
January	15.07 (4.93)	16.38 (4.61)	-1.31
February	18.57 (6.75)	17.23 (4.55)	1.34
March	18.68 (5.26)	16.64 (4.98)	2.04

Note. MI Write Overall Score range = 6.0–30.0.

Figure 1

Percentage of Students that Met their Overall Score Goal Each Month



Calibration

Students set overall performance goals on their monthly GSS before completing each essay. The difference between their goal score and actual score is their calibration accuracy (see Hacker et al., 2008). Hence, we calculated calibration accuracy scores by subtracting students' monthly Overall Score goals from the actual MI Write Overall Scores they achieved each month. Calibration scores closer to 0 represented greater calibration accuracy. We used the absolute value of the calibration score in subsequent analyses.

To determine if student goal calibration scores improved over time, we employed a longitudinal multilevel model with fixed effects and random slopes using the `nlme` package in R (Pinheiro et al., 2021). In this model, goal calibration scores served as the dependent variable and time served as the predictor variable. Time was centered such that November (i.e., baseline) was the intercept. Based on data from students who completed both the GSS and the persuasive essay in a given month (see Table 1), we identified univariate outliers for monthly goal calibration scores grouped by month using the `identify_outliers` function in R (Kassambara, 2021). Of the 136 total goal calibration scores across 56 participants, seven scores from six participants were identified as outliers and removed from the dataset.

We compared a fixed effects model with and without random slopes using the ANOVA function in R (R Core Team, 2021). The model with random slopes demonstrated a significantly better fit. We then explored models with linear and natural logarithmic effects of time. Results from these models are provided in Table 3. AIC and BIC fit statistics indicated that the logarithmic growth model was better fitting than the linear model. Thus, we retained the natural logarithmic growth model when interpreting results (see Figure 2). The significant negative effect of the logarithmic slope suggest that students became more calibrated over time, experiencing rapid improvements in calibration at first that gradually slowed as their accuracy continued to improve.

Figure 2

Growth in Calibration Accuracy Over Time

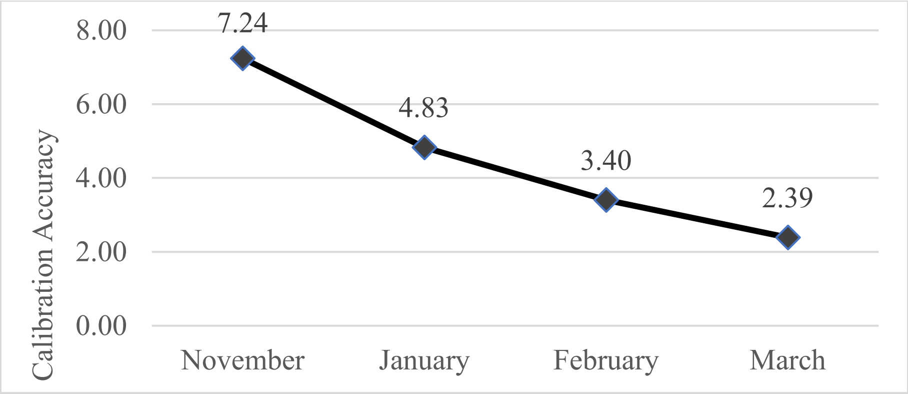


Table 3

Longitudinal Multilevel Growth model for Calibration Accuracy

	Linear Model Coefficient (C.I.) *df = 78*	Logarithmic Model (ln) Coefficient (C.I.) *df = 78*
Fixed Effects		
Calibration Accuracy	6.63*** (5.28, 7.98)	7.24*** (5.68, 8.79)
Linear slope	-1.44*** (-2.04, -0.84)	
Logarithmic slope		-3.49*** (-4.86, -2.12)
Random Effects		
Within-student variance *σ2*	9.67	7.93
Between-student variance *τ00*	10.62	16.04
Random slope variance τ_{11}	1.28	9.24
Random slope-intercept correlation ρ_{01}	-0.99	-1.00
ICC	0.30	0.39
Model Fit		
AIC	706.15	689.48
BIC	723.31	706.64
-2LL	694.15 (*df* = 6)	677.48 (*df* = 6)
R^2		
Marginal	0.16	0.20
Conditional	0.42	0.52

Note. $^{*}p \le .05$; $^{**}p \le .01$; $^{***}p \le .001$. *N* = 129 observations (i.e., calibration scores of students across time) and *n* = 50 groups (i.e., students).

Accepted Manuscript
Version of record at: https://doi.org/10.1080/17501229.2022.2077348

RQ2: Confidence

To answer RQ2, we used descriptive statistics (see Table 4) and visual analysis of box plots (see Figure 3) to examine students' monthly confidence ratings for achieving their writing goals. Descriptive statistics indicate that the average confidence ratings increased between January (60.28) and February (68.03) but remained about equal between February (68.03) and March (68.27). Scores became slightly more stable with a decreasing standard deviation in each subsequent month, although there remained a wide range of confidence scores at each timepoint (see Figure 3).

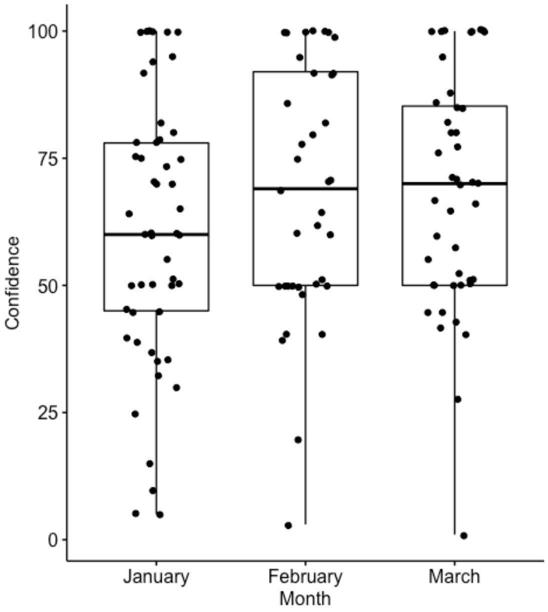
Table 4

Descriptive Statistics for Monthly Student Confidence Ratings

Month	*M*	*SD*	Minimum	Maximum
January	60.28	26.13	5	100
February	68.03	24.95	3	100
March	68.27	23.09	1	100

Figure 3

Boxplot of Student Self-reported Confidence to Achieve Monthly Performance Goals



Based on available monthly GSS data (see Table 1), we employed a longitudinal multilevel model with fixed effects using the `nlme` package in R (Pinheiro et al., 2021) to determine if confidence scores statistically significantly improved over time. In this model, confidence scores served as the dependent variable and time served as the predictor variable. Time was centered such that January represented the intercept. We compared a fixed effects model with and without random slopes using the ANOVA function in R (R Core Team, 2021). Results supported the fixed effects linear model, which indicated a significant and positive effect for time: a one month increase in time was associated with an average increase of 3.36% in students' confidence ratings (95% CI = 0.16% – 6.56%). See Table 5.

Table 5

Longitudinal Multilevel Growth model for Confidence

	Linear Growth Model Coefficient (C.I.) *df = 77*
Fixed Effects	
Intercept	61.34^{***} (54.57, 68.11)
Time	3.36^{*} (0.16, 6.56)
Random Effects	
Within-student variance σ2	236.01
Between-student variance τ00	403.35
Model Fit	
BIC	1183.83
-2LL	1164.40
R^2 (Marginal)	0.033

Note. $^{*}p \leq .05$; $^{**}p \leq .01$; $^{***}p \leq .001$. N = 131 observations (i.e., confidence scores of students across time) and n = 53 groups.

RQ3: Writing Beliefs

Table 6 presents descriptive statistics for writing beliefs at pretest and posttest based on the sample of students with complete pre and posttest data (n = 45)—missing data was due to absences during testing and make-up sessions. Based on a Bonferroni-adjusted alpha level of *p* = .008 (α =.05/6), there were no statistically significant differences between pretest and posttest in overall self-efficacy for writing [$t_{(44)}$ = 1.83, *p* = .073], self-efficacy for idea generation [$t_{(44)}$ = 1.10, *p* = .277], self-efficacy for conventions [$t_{(44)}$ = -0.89, *p* = .380], self-efficacy measured as ability-expectancy beliefs [$t_{(44)}$ = -0.95, *p* = .347], or students' attitudes toward writing [$t_{(44)}$ = -2.08, *p* = .044]. However, there was a statistically significant difference in self-efficacy for self-regulating the writing process [$t_{(44)}$ = 3.43, *p* = .001], with a mean score increase from 64.01% at pretest to 74.46% at posttest.

Table 6

Descriptive Statistics for Writing Beliefs

Belief Construct	Pretest *M* (*SD*)	Posttest *M* (*SD*)
Self-efficacy[a]	69.98 (19.92)	73.83 (17.75)
Conventions	79.87 (19.11)	75.49 (18.32)
Idea generation	69.70 (20.52)	71.60 (20.16)
Self-regulation	64.01 (23.48)	74.46 (19.81)
Ability and expectancy beliefs[b]	2.72 (0.60)	2.57 (0.76)
Attitude toward writing[b]	2.56 (0.89)	2.26 (0.93)

Note. *N* = 45 students with complete data. [a]Range from 0-100 percent; [b]Range from 0 to 4.

RQ4: Writing Performance

Descriptive statistics regarding students' writing performance are presented in Table 7. We employed paired samples *t*-tests with Bonferroni correction (*p* = .025) to evaluate mean differences in their Overall Score from pretest to posttest. Students performed statistically significantly better on their first drafts [$t_{(44)}$ = 3.12, *p* = .003] and their final drafts [$t_{(44)}$ = 3.85 *p* < .001] at posttest.

Table 7

Descriptive Statistics for Pre and Posttest MI Write Overall Scores

	M (*SD*)	Minimum	Maximum
Pretest			
First draft	10.91 (3.20)	6.0	18.7
Final draft	13.25 (3.29)	6.4	20.8
Posttest			
First draft	12.50 (3.69)	6.0	23.2
Final draft	15.40 (4.27)	6.0	24.9

Note. *N* = 45 students with complete data. Range of MI Write Overall Score is 6.0–30.0.

RQ5: Intervention Perceptions

On average, participants agreed that the MI Write was easy to use (*M* = 2.13, *SD* = 0.43), useful for improving their writing ability (*M* = 2.11, *SD* = 0.38), and generally desirable (*M* = 1.79, *SD* = .74) with most students agreeing that they would recommend MI Write to other students (78%) and continue using the program (67%). Students also held positive attitudes toward the GSS, agreeing that the GSS was easy to use (*M* = 2.21, *SD* = .51), useful (*M* = 2.07, *SD* = 0.39), and generally desirable (*M* = 1.90, *SD* = .71) with most students agreeing that they would recommend the GSS to other students (81%) and continue using this survey (72%).

Discussion

Innovations are required to improve students' writing outcomes, especially among populations of students that have historically experienced gaps in opportunity and achievement, such as students from urban school districts, from racial minorities, and who experience poverty. One such innovation is the technology-based pilot intervention described in this study, a first to integrate self-regulation via reflection and goal setting with automated feedback from AWE. The intervention assisted middle-school students—who were Black or Latinx female students living in a high-poverty urban locale—to set and monitor goals related to their writing process, writing product, and overall writing performance. Results indicated that over time students increasingly attained their performance goals and concomitantly increased their calibration accuracy and confidence to attain their goals. Moreover, pretest-posttest comparisons indicated that students increased their self-efficacy for self-regulation and their first-draft and final-draft writing performance. Students also reported positive attitudes regarding the pilot intervention's social validity.

Results suggest that the pilot intervention was associated with positive changes in those outcomes most closely associated with self-regulation for writing (i.e., proximal outcomes like goal attainment, calibration accuracy, confidence to achieve goals, self-efficacy for writing self-regulation, and writing performance). There was no association between the pilot intervention and changes in other writing beliefs that were related to, but less closely linked with, the intervention's focus on self-regulation, such as self-efficacy for idea generation, self-efficacy for writing conventions, overall self-efficacy, self-efficacy measured as ability-expectancy beliefs, or attitudes toward writing. Results highlight the promise of the pilot intervention as well as the need to further develop and strengthen the intervention in the future.

Although we hypothesized that the pilot intervention may influence a measure of self-efficacy that assessed students' broader ability-expectancy beliefs about themselves as writers, we found no gains in this construct from pretest to posttest. One possibility is that the intervention did not include opportunities for students to consider their growth in performance from baseline (pretest). The GSS was structured to facilitate reflecting on the prior month's performance and setting goals for the current month. Thus, it may have been hard for students to apprehend their more comprehensive growth as writers when solely examining incremental month-to-month changes. Another possibility is that the way we measured students' ability-expectancy beliefs masked more subtle changes in these beliefs. Specifically, the measure we employed included items asking students to rate their ability relative to their peers. Students' ability-expectancy beliefs may have shifted but not the sense of their ability relative to their peers. Future research should consider providing opportunities for students to reflect across multiple months of performance and utilize other measures of writing self-efficacy that ask students to compare their current ability relative to their previous ability.

Likewise, students' attitudes toward writing did not improve. One possibility for this finding is that our research design limited students to writing persuasive texts each month. We focused on persuasive writing because of its importance at this level (Ferretti & Graham, 2019), and we elected to keep the writing genre consistent given prior findings of differential performance across genres (Graham et al., 2016). However, such consistency meant that students did not utilize the goal setting survey or MI Write with other genres, such as narrative or creative writing. Another possibility is that limiting students to independent writing without the input from teachers for goal setting or peer collaboration for writing may have limited the effectiveness of the pilot intervention with respect to attitudinal changes. Future research should consider applying our intervention with a greater range of authentic writing experiences and genres (Boscolo & Gelati, 2018), and embedding opportunities for students to experience a collaborative and supportive writing community (see Graham et al., 2012).

Limitations

First, the study employed a single group pretest-posttest design. While such a design is sufficient for investigations at the pilot or initial development stage (e.g., MacArthur & Philippakos, 2013), it is not possible to make causal claims about the impact of our pilot intervention on outcomes of interest. Future research should employ rigorous experimental/control group designs and should evaluate effectiveness within diverse population in terms of gender, school type, and geographical region. Future research also should utilize additional qualitative research methods, such as interviews, to further probe students' perceptions of their self-efficacy, attitudes toward writing, and abilities, and the social validity of the GSS and MI Write.

Second, the study was conducted during school year 2020–2021 amidst the COVID pandemic. This meant that there were stretches of time when school was virtual, and even when school returned to in-person learning, there were students who remained virtual. Consequently, rates of missing data for the monthly goal-setting survey and writing prompt were sizable at times, December in particular. Thus, findings regarding goal attainment, calibration accuracy, and confidence must be interpreted contextually.

Finally, this intervention primarily relied on implicit learning via practice and reflection to foster growth in self-regulation and outcomes of interest. Although implicit learning is effective (Frensch & Rünger, 2003), future research should involve explicit instruction and feedback related to goal setting, as explicit instruction and feedback may further increase learning outcomes (see Graham et al., 2012).

Implications for Future AWE Development

Given the promising findings, AWE developers might consider integrating goal setting supports directly within their systems. To our knowledge, no such functionality exists among AWE systems. Study findings suggest that this may be a beneficial area of future AWE development.

Second, although AWE feedback generally helps students improve their writing quality (Graham et al., 2015; Stevenson & Phakiti, 2014), AWE feedback tends solely to focus on the writing product, leaving feedback on the writing process unaddressed. Thus, another area of AWE development is to provide students with automated feedback on the goals they set. For instance, in our study, if a student scored a 15 on their first essay and a 17 on their second essay, but then set a goal of 30 for their third essay, our goal setting survey did not provide feedback indicating that their goal may be unrealistic. Consider algorithms that identify trends in students' performance across essays and provide students with feedback for goal setting within their zone of proximal development. Such feedback might also be returned to the teacher, so that teachers could identify students in need of additional support related to goal setting and writing. In this way, AWE systems could move beyond simply providing feedback on students' writing product and begin providing feedback on students' writing process with the aim of developing students' self-regulation.

In sum, it is our hope to inspire productive AWE development that builds on AWE's existing affordances, expands its capabilities, and increases its potential to improve the teaching and learning of writing, especially for those most at risk.

References

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*, 5–31. https://doi.org/10.1007/s11092-008-9068-5

Boscolo, P., & Gelati, C. (2018). Motivating writers. In S. Graham, C. A. MacArthur, & M. Hebert (Eds.), *Best practices in writing instruction* (pp. 51–79). New York: The Guilford Press.

Bruning, R., Dempsey, M., Kauffman, D. F., McKim, C., & Zumbrunn, S. (2013). Examining dimensions of self-efficacy for writing. *Journal of educational psychology*, *105*(1), 25.

Bruning, R., & Kauffman, D. (2016). Self-efficacy beliefs and motivation in writing development. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (p. 160–173). The Guilford Press.

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* (Final Report NCEE 2012-4010). Washington, DC: National Center for Education Evaluation and Regional Assistance.

Ferretti, R. P., & Fan, Y. (2016). Argumentative writing. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (2nd ed., pp. 301–315). New York: Guilford Press.

Ferretti, R. P., & Graham, S. (2019). Argumentative writing: Theory, assessment, and instruction. *Reading and Writing, 32*(6), 1345–1357. https://doi.org/10.1007/s11145-019-09950-x

Frensch, P. A., & Rünger, D. (2003). Implicit learning. *Current Directions in Psychological Science, 12*, 13-18. https://doi.org/10.1111/1467-8721.01213

Graham, S., Bollinger, A., Booth Olson, C., D'Aoust, C., MacArthur, C., McCutchen, D., & Olinghouse, N. (2012). *Teaching elementary school students to be effective writers: A practice guide (NCEE 2012-4058).* Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/wwc/publications_reviews.aspx#pubsearch.

Graham, S., Daley, S. G., Aitken, A. A., Harris, K. R., & Robinson, K. H. (2018). Do writing motivational beliefs predict middle school students' writing performance? *Journal of research in reading*, *41*(4), 642-656.

Graham, S., & Harris, K.R. (2018). Evidence-based writing practices: A meta-analysis of existing meta-analyses. In R. Fidalgo, K.R. Harris, & M. Braaksma (Eds.), *Design principles for teaching effective writing: Theoretical and empirical grounded principles* (pp. 13-37). Hershey, PA: Brill.

Graham, S., Hebert, M., & Harris, K. R. (2015). Formative assessment and writing. *Elementary School Journal, 115*, 523-547.

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

Graham, S., MacArthur, C., Schwartz, S., & Page-Voth, V. (1992). Improving the compositions of students with learning disabilities using a strategy involving product and process goal setting. *Exceptional Children, 58*(4), 322–334.

Grimes, D., & Warschauer, M. (2010). Utility in a fallible tool: A multi-site case study of automated writing evaluation. *The Journal of Technology, Learning and Assessment, 8*(6).

Hacker, D. J., Bol, L., & Keener, M. C. (2008). Metacognition in education: A focus on calibration. In Dunlosky, J., & Bjork, R.A. (Eds), *Handbook of metamemory and memory* (pp. 429-455). New York: Taylor Franics Group.

Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In C. M. Levy & S. Randall (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 1-27). Mahwah, NJ: Erlbaum.

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

Kassambara, A. (2021). rstatix: Pipe-Friendly Framework for Basic Statistical Tests. R package version 0.7.0. https://CRAN.R-project.org/package=rstatix

Link, S., Mehrzad, M., & Rahimi, M. (2020). Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement. *Computer Assisted Language Learning,* DOI: 10.1080/09588221.2020.1743323

Lyst, A. M., Gabriel, G., O'Shaughnessy, T. E., Meyers, J., & Meyers, B. (2005). Social validity: Perceptions of check and connect with early literacy support. *Journal of School Psychology, 43*, 197-218.

MacArthur, C. A., & Philippakos, Z. A. (2013). Self-regulated strategy instruction in developmental writing: A design research project. *Community College Review, 41*, 176–195. https://doi.org/10.1177/0091552113484580

MacArthur, C. A., Philippakos, Z. A., & Graham, S. (2016). A multicomponent measure of writing motivation with basic college writers. *Learning Disability Quarterly, 39*, 31–43. https://doi.org/10.1177/0731948715583115

Page, E. B. (2003). *Project Essay Grade: PEG.* In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (p. 43–54). Lawrence Erlbaum Associates Publishers.

Pajares, F. (2003). Self-efficacy beliefs, motivation, and achievement in writing: A review of the literature. *Reading &Writing Quarterly*, *19*(2), 139-158.

Pajares, F., Johnson, M. J., & Usher, E. L. (2007). Sources of writing self-efficacy beliefs of elementary, middle, and high school students. *Research in the Teaching of English*, *42,* 104-120.

Palermo, C., & Thomson, M. M. (2018). Teacher implementation of self-regulated strategy development with an automated writing evaluation system: Effects on the argumentative writing performance of middle school students. *Contemporary Educational Psychology, 54*, 255–270.

Palermo, C., & Wilson, J. (2020). Implementing automated writing evaluation in different instructional contexts: A mixed-methods study. *Journal of Writing Research 12*(1), 63-108. https://doi.org/10.17239/jowr-2020.12.01.04

Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team (2021). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-153, https://CRAN.R-project.org/package=nlme.

R Core Team (2021). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/

Reid, R., Lienemann, T., & Hagaman, J. (2013). *Strategy Instruction for Students with Learning Disabilities (2nd ed.).* New York: Guilford Press.

Roscoe, R. D., Allen, L. K., Johnson, A. C., & McNamara, D. S. (2018). Automated writing instruction and feedback: Instructional mode, attitudes, and revising. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 2089–2093. Retrieved from https://journals.sagepub.com/doi/https:// doi. org/ 10. 1177/ 15419 31218 621471

Roscoe, R. D., Allen, L. K., Weston, J. L., Crossley, S. A., & McNamara, D. S. (2014). The Writing Pal intelligent tutoring system: Usability testing and development. *Computers and Composition, 34*, 39–59. http://dx.doi.org/10.1016/j.compcom.2014.09.002

Roscoe, R. D., & McNamara, D. S. (2013). Writing Pal: Feasibility of an intelligent writing strategy tutor in the high school classroom. *Journal of Educational Psychology, 105*(4), 1010–1025. doi:10.1037/a0032340.

Rutherford, T. (2017). Within and between person associations of calibration and achievement. *Contemporary Educational Psychology, 49*, 226-237. https://doi.org/10.1016/j.cedpsych.2017.03.001

Schunk, D. H., & Pajares, F. (2002). The development of academic self-efficacy. In A. Wigfield & J. Eccles (Eds.), Development of achievement motivation (pp. 15–31). San Diego, CA: Academic Press

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.

Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. *Assessing Writing, 19*, 51–65.

Strobl, C., Ailhaud, E., Benetos, K., Devitt, A., Kruse, O., Proske, A., & Rapp, C. (2019). Digital support for academic writing: A review of technologies and pedagogies. *Computers & Education, 131*, 33-48.

Torrance, M., Fidalgo, R., & Robledo, P. (2015). Do sixth-grade writers need process strategies? *British Journal of Educational Psychology, 85*, 91–112. https://doi.org/10.1111/bjep.12065

Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology*, *25*, 68–81. https://doi.org/10.1006/ceps.1999.1015

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.apa.org/doi/10.1037/edu0000311

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

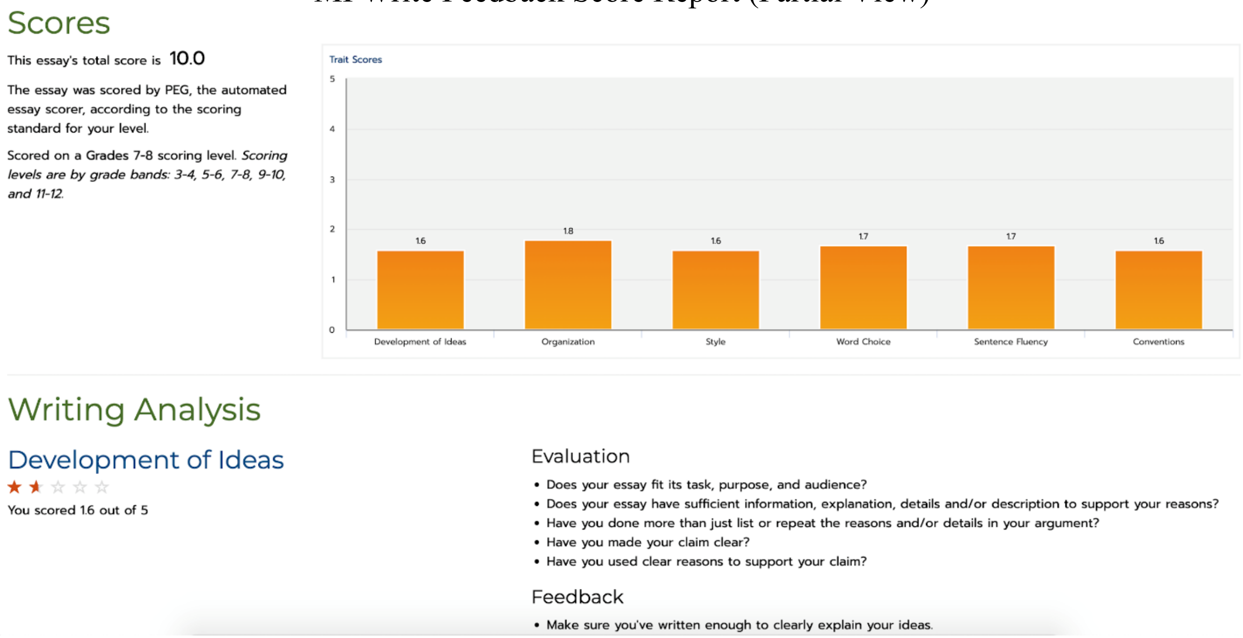
Wilson, J., & Roscoe, R. D. (2020). Automated writing evaluation and feedback: Multiple metrics of efficacy. *Journal of Educational Computing Research, 58*, 87-125. https://doi.org/10.1177%2F0735633119830764

Wolf, M. M. (1978). Social validity: The case for subjective measurement or how applied behavioral analysis is finding its heart. *Journal of Applied Behavior Analysis, 11*, 203–214.

Zimmerman, B. J., & Risemberg, R. (1997). Becoming a self-regulated writer: A social cognitive perspective. *Contemporary educational psychology*, *22*, 73-101.

Appendix A

MI Write Feedback Score Report (Partial View)



Scores

This essay's total score is 10.0

The essay was scored by PEG, the automated essay scorer, according to the scoring standard for your level.

Scored on a **Grades 7-8** scoring level. *Scoring levels are by grade bands: 3-4, 5-6, 7-8, 9-10, and 11-12.*

Writing Analysis

Development of Ideas

★★☆☆☆

You scored 1.6 out of 5

Evaluation

- Does your essay fit its task, purpose, and audience?
- Does your essay have sufficient information, explanation, details and/or description to support your reasons?
- Have you done more than just list or repeat the reasons and/or details in your argument?
- Have you made your claim clear?
- Have you used clear reasons to support your claim?

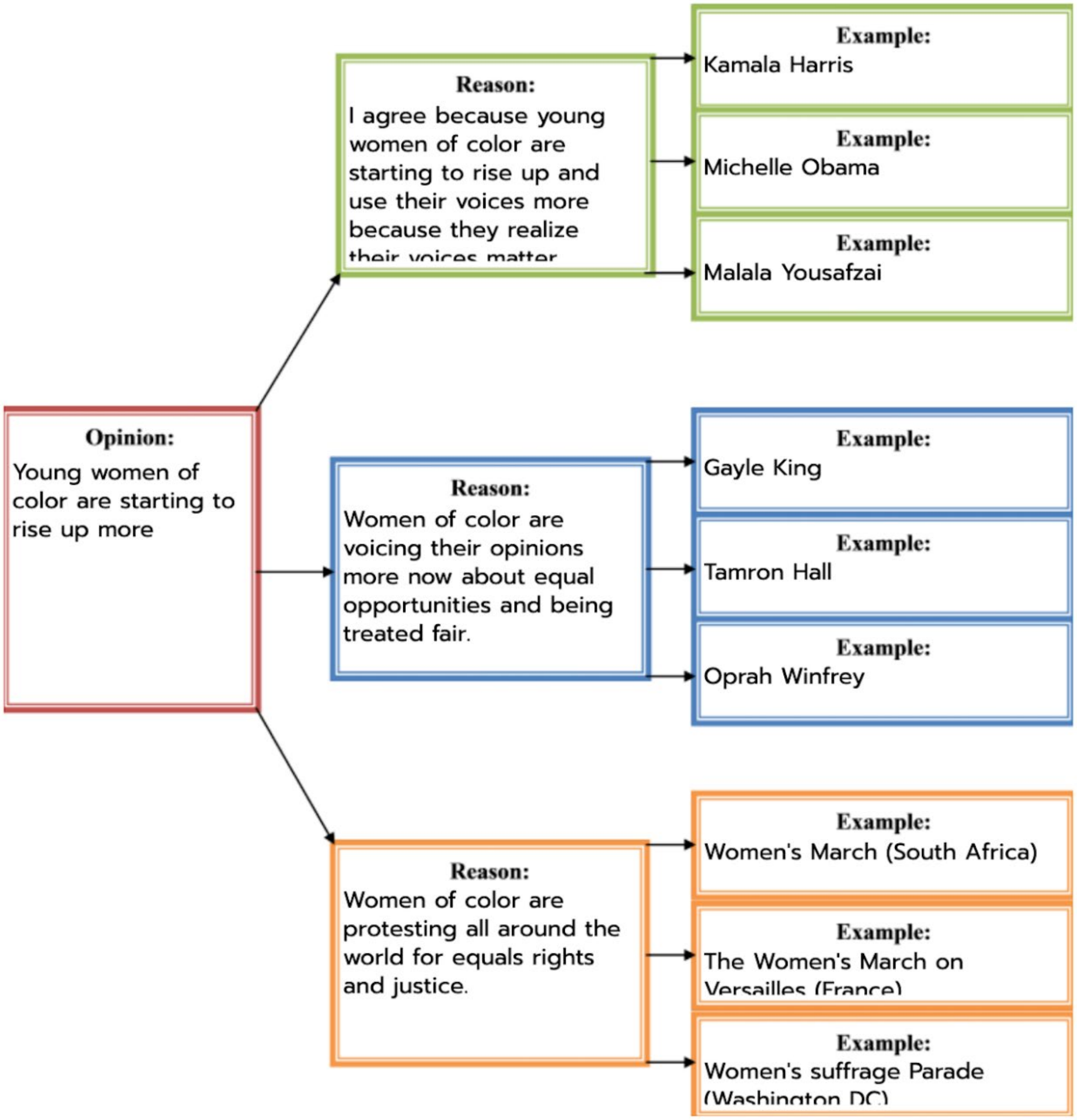
Feedback

- Make sure you've written enough to clearly explain your ideas.

Figure 1. This view illustrates the MI Write automated six-trait scoring along with trait-specific feedback for the "Development of Ideas" trait. The full score report also presents the student with similar "evaluation" and "feedback" information for the remaining five traits.

APPENDIX B

MI Write Argumentative Graphic Organizer



Opinion:
Young women of color are starting to rise up more

Reason:
I agree because young women of color are starting to rise up and use their voices more because they realize their voices matter

Example:
Kamala Harris

Example:
Michelle Obama

Example:
Malala Yousafzai

Reason:
Women of color are voicing their opinions more now about equal opportunities and being treated fair.

Example:
Gayle King

Example:
Tamron Hall

Example:
Oprah Winfrey

Reason:
Women of color are protesting all around the world for equals rights and justice.

Example:
Women's March (South Africa)

Example:
The Women's March on Versailles (France)

Example:
Women's suffrage Parade (Washington DC)

Figure 1. This is an image of one student's completed graphic organizer for an argumentative essay. MI Write includes over 30 digital graphic organizers; this graphic organizer is called the "Argumentative Writing Map."