Automated Feedback and Automated Scoring in the Elementary Grades: Usage, Attitudes, and Associations with Writing Outcomes in a Districtwide Implementation of MI Write

Joshua Wilson[a] (ORCID#: 0000-0002-7192-3510)

University of Delaware

Yue Huang[b]

University of Delaware

Corey Palermo[c] (ORCID#: 0000-0003-1921-5127)

Measurement Incorporated

Gaysha Beard[d]

Red Clay Consolidated School District

Charles A. MacArthur[e] (ORCID#: 0000-0002-7202-5536)

University of Delaware

Author Note

[a]Joshua Wilson, Ph.D., School of Education, University of Delaware; [b]Yue Huang, M.Ed., School of Education, University of Delaware; [c]Corey Palermo, Ph.D., Measurement Incorporated; [d]Gaysha Beard, Ed.D., Red Clay Consolidated School District; [e]Charles A. MacArthur, Ph.D., School of Education, University of Delaware.

This research was supported by Grant R305H170046 from the Institute of Education Sciences, U.S. Department of Education, to the University of Delaware. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education, and no official endorsement by these agencies should be inferred. 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., 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*, 324-276. https://doi.org/10.1007/s40593-020-00236-w

Abstract

This study examined a naturalistic, districtwide implementation of an automated writing evaluation (AWE) software program called *MI Write* in elementary schools. We specifically examined the degree to which aspects of MI Write were implemented, teacher and student attitudes towards MI Write, and whether MI Write usage along with other predictors like demographics and writing self-efficacy explained variability in students' performance on a proximal and distal measure of writing performance. The participants included 1935 students in Grades 3–5 and 135 writing teachers from 14 elementary schools in a mid-Atlantic school district. Findings indicated that though MI Write was somewhat under-utilized, teachers and students held positive attitudes towards the AWE system. Usage of MI Write had a mixed and limited predictive effect on outcomes: The number of essays written had a small predictive effect on state test performance for Grades 3 and 5; gain on revision had a moderate predictive effect on posttest writing quality and a small predictive effect for Grade 5 state test performance. Students' average AWE scores showed consistently moderate to large predictive effects for all outcomes. Interpreted in light of the underlying architecture of MI Write, findings have implications for other school districts considering implementing AWE as well as the design of AWE systems intended to support the teaching and learning of writing.

Keywords: automated feedback, automated writing evaluation, automated essay scoring, writing assessment, educational technology

Automated Feedback and Automated Scoring in the Elementary Grades: Usage, Attitudes, and Associations with Writing Outcomes in a Districtwide Implementation of MI Write

Despite the importance of writing as a critical 21st century skill (National Commission on Writing [NCW] 2003, 2004, 2005), roughly two-thirds of students in the United States in grades four, eight, and twelve fail to achieve grade-level writing proficiency (National Center for Education Statistics 2011; Persky et al. 2002; Salahu-Din et al. 2008). Consequently, students with weak writing skills are at increased risk of referral to special education and school dropout, and of failure to secure stable and gainful employment (Graham and Perin 2007). In turn, the societal costs of remediating weak writing skills are high. Postsecondary institutions, and their students, must assume the costs associated with providing non-credit remedial writing courses. Likewise, private corporations and state governments assume the costs associated with poor writing, costs estimated as billions of dollars annually (Bernoff 2017; NCW 2004, 2005).

Changing writing outcomes in the United States requires changing writing instruction. Indeed, national surveys indicate that students routinely lack opportunities to practice and receive feedback on their writing (Applebee and Langer 2009; Gilbert and Graham 2010; Kiuhara et al. 2009). This is problematic because writing is a complex skill that develops gradually through interactions with members of writing communities (e.g., teachers, peers, or mentors) (Graham 2019), and feedback is a principal form of social interaction that influences writing development (Biber et al. 2011; Kellogg and Whiteford, 2009; Graham et al. 2012; Graham and Perin, 2007). Feedback is information that indicates (a) an individual's performance relative to a learning goal and (b) ways of improving performance to better reach that goal (Black and Wiliam 2009; Hattie and Timperley 2007). Feedback comes from an agent, such as a teacher, mentor, peer, or technology.

Unfortunately, teachers face a number of barriers with respect to increasing the amount of writing practice and feedback students experience. A primary barrier is the time-costs of evaluating and providing feedback on student writing (Warschauer and Grimes 2008). These time-costs make it nearly impossible to provide timely feedback, a characteristic of effective feedback (Dujinhower et al. 2010; Pajares 2003; Shute 2008).

Moreover, providing useful instructional feedback is difficult. Writing incorporates multiple low-level skills, such as spelling, punctuation, capitalization, grammar, and sentence structure; as well as high-level skills, such as word choice, organization, idea development, and style. Low-level skills support translation and transcription (Hayes 2012), the act of turning thoughts into semantically and syntactically accurate text. High-level skills support the development of ideas in writing. Prioritizing feedback among low and high-level skills can be difficult for teachers (Parr and Timperley 2010). Teachers often prioritize feedback on low-level skills, which in turn does little to help students improve their overall writing quality (Clare et al. 2000; Matsumara et al. 2002). Unless the barriers associated with the time-costs and challenges of evaluating writing are addressed, students are unlikely to experience sufficient amounts of practice and feedback needed to improve as writers.

One means of addressing these barriers is the use of automated writing evaluation (AWE). Though the field is absent an agreed-upon, highly specific definition of AWE and of the nature of automated feedback, we adopt the definition offered by Warschauer and Grimes (2008) in their seminal paper on AWE—a definition later used by Stevenson and Phakiti (2014) in their review of the effectiveness of AWE for improving writing quality—AWE are computer systems that use "artificial intelligence (AI) to score and respond to student essays" (p. 22). Thus, an AWE system

is one that provides students both with immediate essay ratings and immediate automated feedback to help students improve their writing when revising.

The nature of automated feedback differs between AWE systems, but most often, automated feedback comes in the form of performance and task feedback, rather than process feedback or self-regulatory feedback (see Hattie and Timperley 2007), though there has been recent work to expand AWE feedback to include process feedback (Deane et al. accepted; Vandermeulen et al. 2020). Specifically, AWE feedback most often takes the form of error correction (e.g., grammar and spelling correction) and suggestions for improving writing quality when revising (i.e., "feed forward"). Some AWE systems' feedback focuses on specific aspects of writing quality, called traits (e.g., the *MyAccess* and *MI Write* AWE systems), while other systems' feedback focus on key elements within an essay such as a thesis or topic sentence (e.g., *Revision Assistant*), content accuracy (e.g., *Summary Street*), or provide a holistic score and separate diagnostic feedback related to English conventions, style, and organization and development *(e.g., Criterion)*.

In addition to automated feedback, AWE systems often include other affordances such as an electronic portfolio to assist students in monitoring personal writing growth, learning management functions to assist teachers in monitoring students' progress (e.g., reporting features), commenting functions to facilitate teachers providing feedback to students (e.g., in-line and messaging features), peer review functionality (either anonymous or identifiable peer review), and embedded interactive skill-building lessons. These affordances are intended to facilitate pedagogically-useful interactions between the AWE system, students, teachers, and peers.

Though AWE's affordances are seen by many as having the potential to transform instruction by vastly accelerating the practice-feedback loop (e.g., Kellogg et al. 2010), AWE is

subject to a number of critiques that chiefly concern the consequences of AWE usage. For instance, there have been concerns that AWE will replace the teacher as the primary agent of feedback (Ericcson and Haswell 2006; Herrington and Moran 2001) and undermine the inherently social nature of writing (National Council of Teachers of English [NCTE] 2013). In addition, concerns over the susceptibility of automated scoring systems to gaming behavior (Bejar et al. 2014; Higgins and Heilman 2014) fuel critiques that AWE narrows and misrepresents the writing construct (Perelman 2014). As a result of such concerns, some groups have rejected the use of AWE (Conference on College Composition and Communication 2014; NCTE 2013).

Despite such criticisms AWE's prevalence in K–12 education in the United States has continued to increase. However, there has been very little research focusing on naturalistic implementations of AWE in school settings, and even less research focused on AWE implementation in the elementary grades. As such, important questions remain regarding: the degree to which AWE is utilized when implemented in the elementary grades, elementary teachers' and students' attitudes towards AWE, and whether AWE usage is associated with elementary grade students' improved performance on proximal and distal outcomes.

Implementation of AWE

Research pertaining to classroom implementation of AWE is a small but growing area of inquiry (Stevenson 2016; Wilson and Czik 2016). Absent of research on AWE implementation in the elementary grades, findings from studies focusing on older students indicate that teachers tend to underutilize and under-implement AWE. Indeed, studies have shown that students often use AWE to submit texts only once, rather than using AWE to support revision (Attali 2004; Shermis et al. 2004; Warschauer and Grimes 2008). In addition, AWE usage tends to be intermittent during the school year, and some features are used inconsistently or not at all (Roscoe and McNamara,

2013). Other studies report low AWE implementation rates; for example, students completing an average of 2.3 essays within a school year (Warschauer and Grimes 2008), or 3–4 essays that were revised an average of 3–4 times each (Foltz et al. 2013; Grimes and Warschauer 2010). Factors cited as barriers to AWE implementation are the availability of sufficient computing resources, the pressure teachers face to keep pace with a curriculum that de-emphasizes writing (Warschauer and Grimes 2008; Wilson and Roscoe 2019), the inflexibility of certain AWE programs for accommodating customized or curriculum-specific prompts (Warschauer and Grimes 2008), and lack of coordinated district-level implementation plans (see Mayfield and Butler 2018).

Attitudes Toward AWE

Teachers generally hold positive attitudes toward AWE while also noting its limitations (Klobucar et al. 2013; Palermo and Thomson 2018; Roscoe and McNamara 2013; Stevenson 2016; Warschauer and Grimes 2008). In a study of three school districts' implementation of an AWE system called *MyAccess* in the middle grades, Grimes and Warschauer (2010) found teachers held very positive attitudes regarding the system's ability to save teachers time, to help students be more motivated to write, and to make writing easier. Teachers also (a) agreed that using AWE simplified classroom management since students were more autonomous, and (b) perceived AWE to be effective for a diverse group of students including general education students, English language learners (ELLs), students with disabilities, gifted students, and at-risk students. At the same time, teachers were less confident in the accuracy of the automated scoring and the ability of AWE to make teaching more enjoyable or to shift teachers' focus to higher-level concerns of writing. Comparable conclusions were reached in studies of middle school teachers' use of an AWE system called *PEG Writing* (now called *MI Write*). Teachers were asked to utilize AWE for half their course sections and GoogleDocs for the other half. When comparing the systems,

teachers rated the AWE system as easier to use and more efficient, and as promoting greater student independence, greater student motivation, greater writing quality, and allowing teachers to devote more of their own feedback to higher-level writing skills (Wilson and Czik 2016; Wilson and Roscoe 2019).

English-speaking students have tended to be less skeptical and critical than teachers, and have expressed positive views towards AWE (Palermo and Thomson 2018), even regarding its scoring and feedback capabilities (Klobucar et al. 2013)—it is important to note, however, that college-aged English learners, particularly in English as a Second Language and English as a Foreign Language contexts have been critical of AWE's error correction and feedback capabilities (e.g., Bai and Hu 2016; Chen and Cheng 2008). Middle-school students using MyAccess reported it to be easy and enjoyable to use, and they perceived its suggestions and scores favorably. They also reported that using MyAccess helped boost their confidence as writers. High school students using an AWE system called *Writing Pal* reported favorable opinions with respect to its ease of use, relevance, understandability, and feedback accuracy (Roscoe et al. 2018). To our knowledge, there have been no prior studies of elementary teachers' and students' attitudes towards AWE.

Outcomes of AWE Usage

Studies sampling students in Grades 3–12 have documented positive effects of AWE usage on improvements in overall writing quality when revising with the aid of automated feedback (Wilson 2017; Wilson and Andrada 2016; Wilson et al. 2014; Foltz et al. 2011; Stevenson and Phakiti 2014). AWE systems also support improvements in nuanced writing and writing-related skills such as increasing the number of explicit citations (Britt et al. 2004), improving the quality of text evidence used in essays (Zhang et al. 2019), increasing the degree to which peer feedback identifies the location of the problem in a text (i.e., peer feedback localization; Nguyen et al. 2017),

and motivating productive revisions of science explanations (Tansomboon et al. 2017). Though less common, studies have also documented positive transfer effects, that is, improvements in the quality of students' writing without the aid of automated feedback (Caccamise et al. 2007; Franzke et al. 2005; Palermo and Thomson 2018). AWE may support writing outcomes by helping students internalize feedback through repeated practice and exposure (see Zellermayer et al. 1991) and to help students calibrate their performance against standards of writing quality, thereby facilitating motivation and goal setting processes aimed at improvement (Moore and MacArthur 2016).

Findings are equivocal regarding the relationship between AWE usage and distal outcomes such as performance on state test of writing and English language arts (ELA). Shermis et al. (2004) found no effect on state test writing performance for tenth-graders assigned within classrooms to either an AWE condition or a business-as-usual condition. Alternatively, Wilson and Roscoe (2019) found that sixth-grade students assigned by classroom to use AWE outperformed peers in classrooms assigned to use GoogleDocs on a state test of ELA; this effect was partially mediated by improvements in writing self-efficacy for students in the AWE condition. One potential explanation for these contrasting findings is the method of AWE implementation. The use of within-class assignment in the Shermis et al. (2004) study likely prevented the kind of broader instructional integration of AWE necessary to influence performance on distal outcomes, such as assigning more writing, increasing the amount of student revision, and changing teacher feedback (Wilson and Czik 2016; Wilson and Roscoe 2019). To our knowledge, no prior research has examined associations between AWE usage and distal outcomes for elementary students.

MI Write

The present study investigates effects on teaching and learning associated with a districtwide implementation of the AWE system *MI Write* (www.miwrite.net), which is developed

and marketed by Measurement Incorporated. MI Write is a web-based AWE system intended to facilitate the teaching and learning of writing by providing students with automated scores and feedback. Prior to Fall 2019, MI Write was known as *PEG Writing*; however, for simplicity, the system will be referred throughout the paper by its current designation, MI Write.

Architecture. Like other AWE systems, MI Write's architecture is designed to mimic the ratings and feedback a teacher would provide to a student's essay. This requires analyzing a representative sample of training essays and constructing models that map text features to scores while minimizing error with human scores assigned to the training essays. Separate models are constructed for each of six traits of writing quality (see Features below). Generalization performance is assessed using a test set of responses held-out from training (see Williamson et al. 2012 for evaluation criteria). MI Write's architecture has been informed by insights from the fields of computational linguistics, machine learning, and natural language processing (NLP). MI Write is powered by a scoring engine known as *Project Essay Grade* (PEG), which performs feature extraction of over 800 linguistic variables theoretically and empirically related to writing quality.

In order to provide ratings and feedback for both pre-packaged and custom prompts, scoring models are created for six traits of writing in three different genres (i.e., narrative, argumentative, and informative/explanatory) and five different grade bands (i.e., 3–4, 5–6, 7–8, 9–10, 11–12) combination, rather than for specific prompts. Thus, the training sample for each model includes responses to a wide range of prompts within a given genre and grade band. A monotonicity constraint across grade bands is built into the models to reflect a general increase in writing quality from one grade to the next. These constraints ensure that if identical essays are submitted associated with two different grade bands, the score for a given genre/trait combination at the higher grade band can be no higher than the corresponding lower grade band score. This

gives rise to a step-wise training approach for each genre/trait model across grade bands: first train the central grade band model using only non-negativity constraints to enforce the positive/negative hidden node group dichotomy, then proceed out from the central grade band, augmenting the non-negativity constraints with bound constraints relative to the network weights of the adjacent grade band model just trained. This method enables PEG to yield scores that are highly consistent with those a human rater would assign: quadratic weighted kappa values across trait × genre × grade-band models average in the low .800's (C. Palermo, personal communication, April 24, 2017).

However, scoring accuracy is not the sole goal of MI Write. Interpretability is an equally important goal, and a goal that is reflected in MI Write's architecture. In order to emphasize the goal of supporting improvements in students' writing quality across drafts, scoring models are designed to be tied directly to engine feedback while maintaining sufficient flexibility to approximate the human scoring process. As such, PEG relies on a modest set of instructionally-relevant (i.e., explicit and easily interpretable) features for which meaningful feedback statements can be provided as inputs to a specially constructed and constrained neural network.

The well-known universal approximation theorem ensures that a neural network with sufficient hidden nodes (i.e., artificial neurons) can approximate a complex scoring function with arbitrary accuracy (Hornick 1991). Compared to a network consisting of only input and output nodes, networks with hidden nodes (which group and transform input features prior to mapping to output scores) support input-output relationships of greater complexity. To maintain interpretability and intimate connection to the generated feedback, within PEG each node in the single hidden layer is specifically designed so that each instructionally-relevant feature enters exactly one hidden node. In this way, each hidden node defines a group of features specifically chosen to be homogenous with respect to how it should affect the trait score (positively or

negatively). Writing experts draw on their understanding of the constructs of each of the six traits of writing quality in selecting the hidden node groups separately for each trait, such that the grouped features measure a common component of writing quality that can be described in the feedback. One such group may measure vocabulary sophistication using count of sophisticated words, count of technical words, and count of hypernyms; another group may measure organization using count of transition words, count of logical adverbs, and cohesion and coherence measures. Each group can thus be interpreted as measuring a different aspect of writing.

This architecture allows the model weights to be constrained in such a way that if a student revises an essay to increase the presence of a positive feature, or decrease the presence of a negative feature, his/her score will increase. By considering the gradient of the scoring network evaluated at the student's response, PEG prioritizes those feedback features most likely to produce the greatest gains in writing quality if altered. The sigmoidal activation function applied to each hidden node output naturally encourages students to attend to feedback across *all* quality aspects because the relative contribution each aspect makes to the overall trait score saturates. This means that once a student attends to feedback that increases the contribution of a given aspect to the point of saturation (i.e., negligible gradient), subsequent feedback will relate to other aspects exclusively. In this way, MI Write's architecture attempts to reflect pedagogical best practice by encouraging students to revise in ways that improve the quality of the whole text (see Fitzgerald 1987; MacArthur 2016) and steer students away from the typical and ineffective strategy of revising a single aspect of the text, typically word or phrase level changes, while ignoring sentence or text level changes (e.g., Chappelle et al. 2015; Kellogg et al. 2010).

Features. Within MI Write, PEG provides students with quantitative feedback in the form of a trait score (range = 1–5) for each of six traits of writing quality and a holistic score formed as

the sum of the trait scores (range = 6–30). The six traits are development of ideas, organization, style, sentence fluency, word choice, and conventions; traits are based off of the widely used Six Trait Scoring model (Northwest Regional Educational Laboratory 2004). The purpose of the quantitative feedback is to help students calibrate their performance relative to grade band and genre-specific rubrics of writing quality.

Students also receive qualitative feedback on these six traits. For instance, an example of feedback for the organization trait is, “Be sure to bring your essay to a close with a good conclusion,” an example of feedback for the development of ideas trait is, “Although your story is well developed, think about whether you can add even more details to improve your story,” an example of feedback for the style trait is “Try making your writing more lively by using more colorful language, dialogue, and questions,” and an example of feedback for the sentence fluency trait is, “Many of your sentences are very simple and begin in the same way. Add introductions or descriptive phrases to add more information and create more interesting sentences.” Students receive one to five feedback statements for each trait. Lastly, students receive feedback in the form of spelling and grammar suggestions, which are presented to the student by highlighting segments of the text with potential errors.

A unique feature of MI Write is its ability to provide quantitative and qualitative feedback on either pre-packaged or customized writing prompts. These prompts can include various types of stimulus materials such as texts, websites, videos, or images. Teachers can also require students to utilize MI Write’s prewriting function by assigning electronic graphic organizers.

In addition to these functions, MI Write has a user interface that serves as an interactive learning environment, supporting a number of pedagogically-relevant interactions between the system, student, peer, and teacher. System-student interactions consist of students receiving

automated scoring and automated feedback, as described above. MI Write also provides students with customized suggestions directing students to complete certain of MI Write's multimedia, interactive skill-building lessons. Finally, students can reference individual electronic portfolios to review their writing and their growth across the school year.

Student-peer interactions consist of peer review, which can be conducted in an identifiable or anonymous manner—MI Write facilitates both single- or double-blind review. Teacher-student interactions consist of students receiving feedback from their teachers via typed comments embedded within the text of their essay and summary comments delivered at the bottom of students' score reports. Finally, teacher-system interactions consist of teachers using MI Write's reporting and learning management functions to monitor student progress and identify classwide or individual learning needs.

Features are intended to support improvements in students' writing behaviors, writing knowledge, and motivation and attitudes towards writing. Individually as well as collectively, these features are theorized to afford improved writing performance and proficiency via:

- *Prompts/practice interface*: increased opportunities for writing practice
- *Electronic graphic organizers*: increased planning behaviors
- *Automated trait-specific feedback that saturates*: increased substantive/comprehensive revisions
- *Automated grammar and spelling correction*: increased editing behavior
- *Automated scoring*: increased calibration accuracy, motivation, and self-efficacy
- *Score reporting across time*: increased motivation and self-regulation by aiding self-monitoring of progress

- *Anonymous or identifiable peer review*: increased peer review opportunities, which in turn, should prompt greater substantive/comprehensive revisions
- *Interactive skill-building lessons*: increased writing knowledge

Present Study

Though prior research has shown generally positive attitudes towards AWE as well as positive effects of AWE on students' writing performance, there is a dearth of research on large-scale, district or statewide implementations of AWE (c.f., Grimes and Warschauer 2010; Warschauer and Grimes 2008). Furthermore, prior AWE research has virtually ignored implementation at the elementary level. In absence of such research, school districts considering adoption of AWE may be misled by effects of researcher-led or laboratory studies of educational technology that may not transfer to naturalistic implementations where participants may not implement, or implement as intended, the educational technology (Newman et al. 2018).

Therefore, the present study was conducted as part of a research-practice partnership (RPP) between a university and a school district to better inform stakeholders of the effects of AWE in a large-scale, naturalistic implementation in elementary schools. A secondary aim of the study was to gather validity evidence to evaluate the theoretical and pedagogical design decisions enacted within the architecture of MI Write. Given this architecture, we hypothesized that students who used more effective strategies and revised across word, sentence, and discourse levels of the text (i.e., across traits) would show gains on measures of writing achievement. Examining correlations between revision outcomes in MI Write and external outcomes would evaluate this premise. Accordingly, the purpose of the present study, which analyzed outcomes after the first year of the partnership, was to answer the following research questions:

RQ1: How and to what extent was AWE implemented in district elementary classrooms?

RQ2: What were Grade 3–5 teachers' and students' attitudes towards AWE?

RQ3: Was AWE usage positively associated with proximal and distal measures of writing performance, namely improvements in writing quality and gains in state test writing performance?

Methods

Context and Participants

In SY 2017–18, a school district in the mid-Atlantic region of the United States implemented MI Write in Grades 3–5 in all 14 of its elementary schools. Based on positive reports from smaller, researcher-led pilots of MI Write in district elementary schools in SY 2014–15 and SY 2015–16, the district elected to implement MI Write districtwide in Grades 3–5 in SY 2017–18. The district implemented MI Write to align to its 1:1 laptop adoption in those grades. In addition, like other school districts that found decreasing writing instruction and writing outcomes associated with No Child Left Behind's emphasis on reading and mathematics over writing, the district elected to implement MI Write to support efforts at increasing writing practice, particularly writing which requires the extended writing process, and at increasing the amount of instructional feedback students receive.

A total of 3000 students and 175 teachers in Grades 3–5 from all 14 elementary schools participated in the districtwide implementation of MI Write during SY 2017–18. The present study reports results based on the sample of students and their teachers who had complete data on (a) both outcomes of interest, (b) both pretest measures, and (c) the AWE usage variables. Our analytic strategy and choice of analysis software (see Data Analysis section) resulted in listwise deletion of cases with missing data from any one of those sources. Thus, the sample with complete data consisted of 1935 students and 135 teachers in Grades 3–5 from all 14 elementary schools. Demographics of this sample are presented in Table 1. Sample demographics were generally

consistent with district demographics across those grade levels. To ensure that our decision to sample students with complete data would not influence results, we also conducted analyses with the sample that included missing data, and results were consistent. Therefore, for ease and simplicity we report results solely for the sample with complete data.

Outcome Variables

Teachers' Attitudes Toward MI Write. In Spring 2018, an electronic survey was sent by the school district to teachers in Grades 3–5 who used MI Write that year. The survey included questions regarding the ways that teachers were trained to use MI Write, how they implemented the system in their classroom, and their attitudes towards the system. Teachers were asked to rate their agreement using a Likert scale for 20 items about MI Write's ease of use and acceptability (7 statements; e.g., "MI Write is easy to use"), effects on student learning (7 statements; e.g., "MI Write helps students improve their writing"), and effects on instruction (6 statements; e.g., "I assign more writing when using MI Write"). Overall reliability (α) for these 20 items was 0.96.

Students' Attitudes Toward MI Write. In Spring 2018, students completed a survey that probed their attitudes towards MI Write via nine items regarding its ease of use (e.g., "I understand how to use MI Write"), desirability (e.g., "I want to use MI Write again next year"), and perceived effectiveness for promoting motivation (e.g., "MI Write helps me feel more motivated to write") and for improving writing performance (e.g., "MI Write helps me become a better writer"). Overall reliability (α) for these nine items was 0.92.

Writing Quality. Three informative writing prompts per grade (9 total) were created by the research team and embedded within MI Write to assess students' writing quality. Prompts directed students to read two short texts (each approximately 190 words) of equivalent Lexile level and to use information from the texts when composing their response. Prompts topics were selected

in partnership with the district and in consideration of students' grade-level ELA curricula; Lexile levels for the stimulus material increased across grades. For instance, a Grade 3 prompt asked students to compare and contrast information from two texts (both 600–700L) about living in rural and suburban areas; a Grade 4 prompt asked students to compare and contrast information from two texts (both 700–800L) about the characteristics of Native American tribes in the Northeast and Southwest United States; and a Grade 5 prompt asked students to compare and contrast information from two texts (both 800–900L) about tropical and temperate rainforests. Prompts were counterbalanced across schools and pretest/posttest administrations to control for prompt effects.

Teachers were instructed to allot several days for students to complete their writing for each prompt, ensuring sufficient time for students to move through each of three writing process steps: planning, drafting, and revising and editing. Teachers used a set of standardized directions to administer the writing prompts and to inform students exactly what to do in each of the three steps. Specifically, the directions instructed students to read the articles, make a plan, compose a draft, and then revise their essay within MI Write based on the automated feedback they received. In accordance with a process-based approach, students' responses to writing prompts were not timed.

The writing quality of students' final drafts was measured by the PEG holistic score. PEG scores are highly reliable (see Shermis 2014; Shermis et al. 2002): quadratic weighted kappa of machine-human agreement averages in the low .800s (M = .835, SD = 0.026). We elected to use the PEG holistic score rather than the trait scores for two reasons. First, the holistic score is more aligned with other broad, general outcome measures of writing. Second, the trait scores were highly correlated (range r = .75–.94), which likely replicated correlations among the human-scored

training data (see Gansle et al. 2006). Further, when subject to a factor analysis, the trait scores loaded on a single factor that explained 82% of trait-score variance, suggesting that the traits provide similar information.

State Test Writing Performance. Along with 12 other states and additional U.S. territories, Delaware administers the Common-Core aligned Smarter Balanced Assessment Consortium (Smarter Balanced) summative ELA test in Grades 3–8. The writing portion of the ELA test requires students in Grades 3–5 to complete six selected-response items and compose one essay as part of a performance task. Smarter Balanced estimates the performance task requires two hours to complete, though the test is untimed. Selected-response items are scored automatically and the performance task is scored by hand for three traits of writing quality: organization/purpose, evidence/elaboration, and conventions. Students' scores on the writing portion of the ELA test are reported to the district in the form of vertically-scaled scores that range from 2000–3000. The 2017–18 ELA test was subject to rigorous validation (Smarter Balanced 2018) and its marginal reliability, defined as one minus the ratio of mean error variance to observed score variance, was .93 for Grades 3, 4, and 5.

Predictor Variables

MI Write Usage Data. We examined several key MI Write usage variables using log data, including the number of essays students completed across the year, the average number of drafts completed per essay, students' average gain score, and the total number of MI Write lesson minutes students completed. Though MI Write enables peer review, this feature was not used in the district; therefore, we did not include peer review usage as a predictor—a likely explanation for the non-use of MI Write's peer review function was lack of training (see Study Procedures section, below).

Number of essays. Repeated, sustained writing practice is essential for developing writing ability (Kellogg and Whiteford 2009). As an index of students' writing practice, we calculated the number of unique essays (i.e., essays on different topics) students completed within MI Write across the school year. In addition, we also calculated a classroom-level version of this measure, defined as the total number of writing tasks assigned by a teacher by which more than half of the students in the class responded. At the classroom level, we refer to the number of essays variable as "Class Number of Assignments." Accounting for number of essays across both the student and teacher levels allowed us to account for the total number of essays students completed as assignments as well as additional essays a student may have completed of his/her own motivation.

Average drafts/essay. In addition to writing practice, writing ability develops when students repeatedly revise their writing with the support of feedback (Graham et al. 2015). To index the degree to which students revised their writing after receiving MI Write's automated feedback, we calculated the average number of drafts a student completed per essay for the school year. We calculated this variable by dividing the total number of essay drafts a student wrote for the entire year by the total number of unique essays a student completed. We also calculated this measure at the teacher-level to account for differences in the ways that teachers utilized MI Write in their classroom (e.g., for primarily single draft writing or for repeated revision).

Average gain score. To differentiate between exposure to automated feedback—indexed via average drafts/essay—and benefits derived from that exposure, we calculated students' average "gain" score within MI Write. We first calculated the average first draft score for all essays submitted across the year and then calculated the average final draft score for those essays, the score received after a student had finished revising in response to MI Write's feedback. The difference between those scores was the average "gain" score, the amount of score points a student,

on average, improved his/her writing when revising. We did not calculate a comparable classroom-level variable as we considered this construct to be more appropriately measured at the student level. To ensure that students' average gain score measured the intended construct and was not simply a proxy for students' general writing ability, we calculated the correlation between the gain score and students' average-first draft writing score. There was no correlation (r = .002, p = .940), affirming the validity of the average gain score measure.

Total MI Write lesson minutes. MI Write contains a suite of interactive, multimedia skill-building lessons, recommended to students based on their writing quality. We reasoned that usage of these lessons was an important aspect of system implementation, and may explain variance in outcomes of interest. We hypothesized that students who completed more lessons might increase their writing skills and have greater motivation to improve their writing relative to students who completed fewer lessons. To account for this, we calculated the total number of minutes students spent utilizing MI Write's lessons. Since at the time of the study MI Write did not allow teachers to assign lessons to students, we did not calculate a comparable classroom-level variable.

Covariates

Prior writing performance. For both writing outcomes—the posttest PEG holistic score and the state writing test score—we accounted for prior writing performance. For the models predicting the posttest PEG holistic score, we accounted for prior performance by including the pretest PEG holistic score in the models, as well as students' average first-draft score in MI Write (described below). For the models predicting performance on the state writing test, we accounted for students' scores on the prior year's state writing test (for Grades 4 and 5 only) and students' average first-draft score in MI Write (Grades 3–5). We could not account for prior state test writing performance for Grade 3 students because state testing does not begin until third grade.

Average first-draft score. As a measure of students' independent writing performance, we calculated students' average PEG holistic score across all first-drafts they wrote for the school year. MI Write does not provide feedback until students submit their writing for evaluation; hence, first-draft performance is a reasonable indicator of independent writing performance. We utilized this variable to account for students' writing performance within our predictive models (i.e., when predicting posttest writing quality and state writing test performance) and to ascertain whether the PEG scoring system indexed writing performance in a way that aligned (i.e., was correlated) with an external measure of writing performance (i.e., the state writing test). We also included in our predictive models a measure accounting for the average first-draft score at the classroom level.

Writing Self-Efficacy. Given that writing self-efficacy, defined as a student's confidence in his/her writing ability (Bruning and Kauffman 2016), is associated with writing behavior such as writing more in and out of school (Troia et al. 2013) and writing achievement (Graham et al. 2007), our research team developed a writing self-efficacy survey. The survey was administered by the district as a GoogleForm. The 21 item survey, based on those developed by Bruning et al. (2013) and Graham et al. (2012), used Likert-like ratings to assess students' writing attitudes (5 items; e.g., "I like to learn how to write"), self-efficacy for writing conventions (4 items; e.g., "I can write complete sentences"), self-efficacy for idea generation (4 items; e.g., "I can think of many ideas for my writing"), self-efficacy for composing in different genres (4 items; e.g., "I can state an opinion and give reasons and evidence in my writing"), and self-efficacy for engaging in different writing processes (4 items; e.g., "I can revise my writing and make it better"). Reliability for the 21-item survey was high: $\alpha = 0.91$. We thus created an overall writing self-efficacy score by averaging students' scores across the 21 items (range = 1–5), with higher scores indicating greater average writing self-efficacy.

Demographics. In addition to accounting for grade level, given well-established demographic trends in writing performance (Troia et al. 2018), we included the following demographics in our analyses as dummy-coded variables: Gender, Race (dummy coded for African American, Asian, and Hispanic with Caucasian as the reference category), special education status, and ELL status. Data regarding individual students' socio-economic status were not available for reporting. Instead, we measured socio-economic status at the school-level as the percent of students within a school receiving free or reduced lunch.

Study Design and Data Source

Since the entire district implemented MI Write in SY 2017–18 as part of the RPP, the current study employed a single-group, pretest-posttest design. In early Fall 2017, teachers received an initial training to use MI Write. Training was delivered by the first author in the form of a two-hour workshop that addressed how to login to MI Write; view and adjust course information; assign pre-packaged writing prompts; create, assign, and share custom (i.e., teacher-created) writing prompts; review and provide feedback on student writing; and review the reporting functions available within MI Write. Due to time constraints, the initial training did not address all features of MI Write (e.g., teachers were not introduced to the peer review functionality). However, the training also discussed some pedagogical best-practices for using AWE, namely to utilize AWE as a tool to support teacher-led writing instruction, not replace it; to provide teacher feedback to supplement the automated feedback; to create and share customized, curriculum-specific writing prompts to better integrate AWE within the curriculum; and to encourage students to write and revise often. The first author provided additional training on request throughout the school year to grade-level teams and also responded to questions of both

technical and pedagogical nature via email. Through these additional trainings and interactions, some teachers learned how to use additional features of MI Write (e.g., peer review).

In late Fall 2017, teachers administered pretest writing prompts within MI Write. Then, during the year, teachers used MI Write to support writing instruction as they saw fit. The district did not require teachers to assign a certain number of writing prompts within MI Write or require that all student writing should be done using MI Write. Instead, the district provided MI Write as a tool to support writing instruction and encouraged, but did not require, teachers to avail themselves of the affordances offered by this tool.

In Spring 2018, teachers administered a posttest writing prompt within MI Write as well as an electronic survey that asked students about their attitudes towards MI Write, their attitudes towards writing more broadly, and their writing self-efficacy. Also in Spring 2018, teachers responded to a survey asking them to report how they utilized, and their attitudes towards, MI Write. The pretest and posttest writing prompts were required by the district, but because they were administered as part of a research project, there was no consequence to teachers or students for not completing these assignments. In Spring 2018, students completed the Smarter Balanced ELA test, used by the state for accountability purposes. In Summer 2018, we created an anonymized dataset by merging district-provided demographic and state test data from Spring 2017 (prior year) and Spring 2018 with MI Write usage data, teachers' and students' survey responses, and students' pretest and posttest writing prompt data.

Data Analysis

To answer RQs 1 and 2, we utilized descriptive statistics to summarize the degree to which aspects of MI Write were implemented and teachers' and students' attitudes towards the AWE system. For RQ3, to investigate whether use of MI Write was associated with superior writing

quality and state-test writing performance, we utilized hierarchical linear modeling (HLM) to account for the nesting of students within classrooms within schools (Raudenbush and Bryk 2002). For each dependent variable (i.e., the posttest PEG holistic score and the Smarter Balanced writing scale score), we initially specified three-level, fully unconditional models to calculate the intra-class correlation (ICC) and determine whether there was sufficient variance to proceed with a three-level model. Only when the ICC is greater than 10% of the total variance in the outcome would the analyst need to consider multilevel methods (Lee 2000). Thus, we modified our models based on the initial ICC.

For the HLM models predicting the posttest PEG holistic score, the results from the initial three-level models showed that school-level variance was less than 10%. Thus, we designed two-level models—students nested within teachers—to predict the posttest PEG holistic score. However, we included school dummy variables as fixed effects (Allison 2009) to control for any school-level influence. We first specified the unconditional random-intercept model with only school dummy variables as fixed effects at level two. Then we added the group-mean centered student-level predictors and covariates at level one. The final conditional model included grand-mean centered teacher-level predictors at level two.

For the HLM models predicting students' Spring 2018 Smarter Balanced writing scale scores, initial three-level exploratory analyses indicated that teacher-level variance was less than 10% of the total variance. Hence, we specified two-level models that considered students nested within schools. For these models, instead of examining associations between teachers' usage of MI Write and state test writing performance, we aggregated the teacher-level predictors by school and added them at level two in our models. Since we did not have the data for third-graders' 2017 state test scores we conducted separate analyses by grade. For each grade, we first specified

unconditional random-intercept models. Then, group-mean centered student-level predictors and covariates were added at level one. The final conditional model included the grand-mean centered aggregated teacher-level predictors and two additional school-level predictors with known associations with state test performance (Troia et al. 2018): the percentage of students within the school receiving free or reduced-priced lunch (%FRL) and the percentage of students in the school who made adequate growth in ELA performance the prior year.

HLM analyses were conducted using HLM 7 software. For all analyses, we calculated the percent of variance explained in the outcome by comparing the variance components of the final conditional model to those of the unconditional model. As a measure of effect size, we calculated standardized coefficients by multiplying the unstandardized coefficient by its standard deviation and then dividing by the standard deviation of the dependent variable (Lorah 2018; Snijders and Bosker 2012). According to Cohen's (1988) interpretation of standardized coefficients, we identify .05 as a small effect, .10 as a medium effect, and .25 as a large effect.

Results

RQ1: AWE Implementation and Utilization

Teacher Usage: Survey Data. Teachers (n = 69) primarily relied on AWE during their writing instructional period, which averaged 35 minutes daily (SD = 9.93 min; range = 10–45 min; Mdn = 30 min). The majority of teachers agreed or strongly agreed (83%) that they felt adequately trained to use MI Write to teach writing. When asked how they used MI Write in their classroom that year, teachers most frequently required students to complete all or most of the curriculum-required writing assignments plus additional writing activities (50% of respondents). The next most common usages were (a) using MI Write only for the writing assignments required in the district curriculum (30%), and (b) completing only the district-mandated pretest and posttest

writing prompts (10%), or (c) using MI Write as a "center" activity, where students could use MI Write as a form of voluntary, supplemental independent practice (10%).

Teachers reported that the most common features of MI Write they used were, in order of frequency: assigning one of MI Write's pre-packaged prompts to students, assigning a teacher-created prompt that did not have stimulus material, reviewing and printing students' score reports, providing written comments to students through MI Write, sharing a prompt within MI Write with grade-level colleagues, assigning a teacher-created prompt that did have stimulus material, using MI Write's messaging functions to respond to a student question or comment, and to use MI Write's peer review features. This list indicates that the most basic features of MI Write (e.g., assigning a pre-packaged writing prompt and reviewing score reports) were used most frequently, whereas the features that required the greatest facility with the system (e.g., assigning a teacher-created prompt that had attached stimulus material or using MI Write's peer review functions) were used least frequently.

Teacher Usage: Log Data. Descriptive statistics of classroom MI Write usage variables are presented in Table 2. On average, teachers assigned 4.44 writing assignments across the school year and had their students revise each assignment 3.70 times. There were no statistically significant grade-level differences in the average number of assignments [$F_{(2, 130)} = 0.95$, $p = .388$]. There was a marginally statistically significant overall effect of grade-level on the number of revisions completed at the classroom level [$F_{(2, 130)} = 3.01$, $p = .053$], with fifth graders revising to a slightly greater degree than third graders ($p = .062$). The average first-draft PEG holistic score at the classroom level was 16.84. There were statistically significant differences across grades for classroom performance [$F_{(2, 130)} = 4.76$, $p = .010$]: third grade classrooms performed lower than

those of fourth grade (p = .056) and fifth grade (p = .016), but there were no differences between the average performance of fourth and fifth grade classrooms (p = 1.000).

Student Usage: Log Data. Figure 1 shows the number of drafts submitted to MI Write across the school year. Across the district's 14 elementary schools, students submitted a total of 78,582 drafts. Consistent with prior research (Roscoe and McNamara 2013), AWE usage was inconsistent. A large spike in usage in the fall was followed by a decline towards the end of the calendar year, fluctuating usage in the first few months of 2018, and a large spike in usage in May that likely coincided with preparation activities for the state test and district posttesting.

Table 2 reports descriptive statistics of key MI Write usage variables at the student level for the full sample and disaggregated by grade. On average, students completed 6.84 assignments within MI Write across the school year and revised those essays an average of 4.56 times. There were statistically significant grade-level differences for both number of essays completed [$F_{(2, 1932)}$ = 54.66, p < .001] and average drafts/essay [$F_{(2, 1932)}$ = 30.78, p < .001]. Fourth graders completed a greater number of essays than fifth graders who completed a greater number of essays than third graders, and fifth graders revised those essays more than both third and fourth graders who revised to an equal degree (all contrasts p < .001).

Students' average first draft score was 15.28 (SD = 3.82) and students' average gain score was 1.85 holistic score points, indicating that the overall sample made approximately half a standard deviation gain in writing quality between first and final drafts, on average. There were statistically significant grade-level differences for students' average first-draft score [$F_{(2, 1932)}$ = 27.57, p < .001] and average gain score [$F_{(2, 1932)}$ = 31.63, p < .001]. For both variables, fifth graders scored higher than fourth graders who scored higher than third graders (all contrasts p < .001).

Students used MI Write's interactive skill-building lessons infrequently. The median was 0 minutes of lesson usage for the sample. Thus, we created dummy variables to represent ranges of lesson usage: 0 min (59% of sample), 1–12 min (16.5% of sample; 12 min = 75th percentile), and ≥ 13 min (24.2% of sample). A slightly higher percentage of fifth-graders completed 1–12 min of lessons than fourth graders (19% versus 13%; $p = .011$), while a slightly higher percentage of fourth graders completed ≥ 13 min of lessons (28% versus 21%; $p = .026$). There were no other statistically significant grade-level differences for lesson usage.

RQ2: Teacher and Student Attitudes Toward AWE

Table 3 shows descriptive statistics of the survey items evaluating teachers' attitudes toward AWE. The median and modal responses for all but two items—both of which related to the accuracy and validity of MI Write's automated scoring—were 4 on a 5-point Likert-like scale (1 = *Definitely not*; 2 = *Probably not*; 3 = *I don't know*; 4 = *Probably yes*; 5 = *Definitely yes*). There were no statistically significant differences in teachers' attitudes across grade levels for any of the items; thus, for parsimony, Table 3 reports descriptive statistics only for the overall sample.

Table 4 presents descriptive statistics of the survey items evaluating students' attitudes toward AWE. Overall, students' attitudes toward AWE were positive, with the mean of all the items ranging from 3.32 to 4.38 ($SD = 1.11$), the median ranging from 4–5, and the mode being 5 for all items. Interestingly, while teachers perceived MI Write to increase student writing, students themselves reported the lowest agreement with the item "MI Write helps me feel more motivated to write" ($M = 3.32$), suggesting perhaps that while students exhibited more energy to improve their writing—as noted by strong agreement with the item about fixing writing—students' broader motivation and attitude towards writing was not influenced by the adoption of AWE. Using a Bonferroni-corrected alpha of $p = .006$, a series of one-way ANOVAs was conducted to examine

grade-level differences for students' responses to each of the nine items. There were statistically significant grade-level differences for Item 3 ("I believe the scores that MI Write gives") and Item 9 ("I want to use MI Write next year"): $F_{(2, 1909)} = 10.58$ and 5.13 with $p < .001$ and $p = .006$, respectively. In each case, third graders displayed greater mean agreement than their peers.

RQ3: Associations Between AWE Usage and Outcomes

Writing Quality. Results of HLM analyses predicting gains in writing quality as measured by the posttest PEG holistic score are shown in Table 5. After accounting for students' demographics, writing self-efficacy, the pretest PEG holistic score ($\beta = 0.28$, a large effect size), and student's average first-draft score ($\beta = 0.26$, a large effect) and the classroom average first-draft score ($\beta = 0.22$, a medium effect size), one AWE usage variable was associated with gains in posttest writing quality: students' average gain score ($\beta = 0.12$, a medium effect size). Students' total number of essays submitted, average drafts per essay, and lesson usage had no significant associations with gains in writing quality, nor did the classroom number of writing assignments or the classroom average drafts/essay. However, average drafts per essay was moderately correlated with students' average gain score in all grades: $r = .661$ in Grade 3, $r = .781$ in Grade 4, and $r = .678$ in Grade 5. Thus, findings suggest that after accounting for the number of revision attempts students made, students who were more successful at revising their essays, as indicated by greater average gain scores, produced higher quality posttest essays. Overall, student-level variables explained 32.33% of the variance in the posttest PEG holistic score and teacher-level predictors explained 8.20% of the variance.

State Test Performance. Results of the HLM models predicting 2018 Grade 3 performance on the writing portion of the Smarter Balanced state test are presented in Table 6. After accounting for students' demographics, writing self-efficacy, and student's average first-

draft score (β = 0.25, a large effect), only one AWE usage variable was associated with state test writing performance in Grade 3: total number of essays submitted in MI Write (β = 0.08, a small effect). None of the teacher-level MI Write usage predictors (aggregated to the school level) were significant predictors. The only school-level predictor that was statistically significant was the percentage of FRL (β = −0.57, a large negative effect). Overall, student-level and school-level predictors explained 25.40% and 83.90%, respectively, of the variance in third-graders' state test writing performance.

Results of the HLM models predicting 2018 Grade 4 performance on the writing portion of the state test are presented in Table 7. After accounting for students' demographics, writing self-efficacy, prior state test writing performance, and students' average first-draft score (β = 0.11, a medium effect), none of the student-level or teacher-level AWE-usage variables (aggregated to the school level) were significant predictors. The only school-level predictor that was statistically significant was the percentage of FRL (β = −0.39, a large negative effect). Overall, student-level and school-level predictors explained 46.19% and 67.56%, respectively, of the variance in fourth-graders' state test writing performance.

Results of the HLM models predicting 2018 Grade 5 performance on the writing portion of the state test are presented in Table 8. After accounting for students' demographics, writing self-efficacy, prior state test writing performance, and students' average first-draft score (β = 0.19, a medium effect), the following AWE-usage variables were associated with gains in state test writing performance: total number of essays submitted in MI Write (β = 0.07, a small effect) and average gain score (β = 0.08, a small effect). None of the teacher-level MI Write usage predictors (aggregated to the school level) were significant predictors. As with the HLM model predicting posttest writing quality, even though the average number of drafts and average gain score were

moderately correlated, the average gain score was a stronger predictor of writing performance on external measures. There was a large negative effect size of school FRL percentage ($\beta = -0.57$). Overall, student-level and school-level predictors explained 53.81% and 90.01%, respectively, of the variance in state test writing performance.

Table 9 presents a summary of findings for RQ3.

Discussion

This study extended prior research on AWE by examining a naturalistic, districtwide implementation of the AWE system MI Write in the elementary grades. Results indicated that AWE usage fluctuated throughout the school year and the lesson and peer review features were effectively unutilized. Teachers and students held positive attitudes towards AWE and there were varying patterns of association between AWE usage and gains in writing quality and state test performance. Of note—given the architecture of MI Write—was the finding that the average gain score, and not the average number of drafts per essay, was predictive of external outcomes. The average first-draft PEG score consistently showed moderate to large predictive relationships with all outcomes. Findings serve as a benchmark for other districts considering adopting AWE systems to support the teaching and learning of writing in the elementary grades. In addition, findings offer a partial validation of the theoretical, pedagogical design decisions undergirding the architecture of MI Write's scoring and feedback systems.

Implementation: Considerations and Challenges

Findings related to implementation confirm prior research that teachers may not utilize AWE continuously across the school year nor utilize all AWE features (Roscoe and McNamara 2013), which is consistent with naturalistic implementations of educational technology more broadly (Newman et al. 2018). Findings also extend prior research by illustrating that the main

functionality of AWE, that of drafting and revising essays, was utilized to a moderate degree. On average, teachers assigned their classes approximately four essays within MI Write and had their students revise four times, a level of usage exceeding that reported in some prior research (e.g., Attali 2004; Shermis et al. 2004; Warschauer and Grimes 2008) and consistent with that reported in others (e.g., Foltz et al. 2013; Grimes and Warschauer 2010). The average usage for students exceeded that of the classroom assignments; on average, students completed approximately 7 essays and 5 revisions per essay, which is the highest implementation within AWE research reported to date. Differences in classroom and student use averages are perhaps explained by two related factors. First, we defined a classroom assignment as one for which at least 50% of students submitted a response. Teachers could also provide writing assignments for enrichment purposes to individual or groups of students. Second, several teachers implemented AWE as a center activity to increase independent practice. Teachers used AWE flexibly within and across classrooms.

Furthermore, students not only revised their essays several times, they increased their writing performance approximately a half standard deviation between first and final drafts, on average. This finding is consistent with prior research illustrating that AWE motivates greater amounts of revision (Wilson and Czik 2016; Grimes and Warschauer 2010) and that students can productively utilize automated feedback to revise (Wilson and Andrada 2016; Wilson et al. 2014; Foltz et al. 2011). This finding is encouraging because elementary students typically revise very little, especially in ways that increase quality (Fitzgerald 1987; MacArthur 2016). When integrated with teacher-led writing instruction, AWE can be expected to increase the amount of revising, and effective revising, that elementary students conduct.

While there was a moderate degree of utilization of AWE, findings beg the question as to why utilization was not more consistent across the school year, particularly since the intention of

AWE is to increase the amount of regular writing practice students experience. On average, teachers assigned an essay within the AWE system every other month—Why not more?

Unlike prior large-scale implementations of AWE (Warschauer and Grimes 2008), the present implementation did not suffer from lack of technology resources (each student had a laptop), lack of district-level support for the integration of AWE within instruction (the present study was conducted within an RPP), sufficient professional development (teachers felt suitably trained), or lack of an AWE system that functions with customized prompts (MI Write allows teachers to create prompts). Still, the pattern of AWE utilization reflects that of *massed practice*, where students engage in less frequent but highly intensive periods of writing activity, rather than *distributed practice*, where students practice writing frequently and consistently. It is the latter form of practice that is essential for supporting retention and generalization of acquired skills to novel contexts (Archer and Hughes 2011), such as performance on distal measures like a state test.

Thus, in absence of commonly-cited barriers to increasing AWE utilization, it is possible that teachers lacked sufficient technological-pedagogical content knowledge (TPACK; Mishra and Koehler 2006) to understand how to enact instructional best-practice that emphasizes frequent practice (Graham et al. 2012) with the affordances of AWE. Another possibility is that teachers possessed sufficient TPACK but curricular pressures inhibited increasing students' writing practice. Prior research has shown that even when teachers report positive attitudes towards AWE the pressure to keep pace with a curriculum that includes limited opportunities for extended writing practice may trump the time-saving affordances of AWE (Wilson and Roscoe 2019). A teacher who is committed to increasing writing practice can do so using AWE—some teachers in the present study assigned 10 writing assignments—but to transform writing outcomes for students, more composing opportunities may need to be embedded at the curricular level. AWE can be

expected to maximize a teacher's time, and professional development and training can augment teachers' TPACK, but curricular barriers to increasing writing practice may need to be addressed as well.

Our findings also suggest that districts may need to take a more hands-on, authoritative approach regarding expectations for AWE implementation. The school district encouraged but did not require teachers to utilize MI Write; there were no consequences, positive or negative, associated with use or non-use of MI Write. We feel that this approach is wise for districts in the stage of initial adoption—indeed, the current study reports results of the first year of district-wide adoption. However, increasing AWE implementation—and ultimately student writing practice and writing outcomes—may, in later years, require districts to exert greater authority with respect to the role that AWE is to play within the curriculum (see Mayfield and Butler 2018).

Attitudes

Findings related to attitudes towards AWE were consistent with prior research showing that teachers tend to hold positive attitudes while also recognizing AWE's limitations (Klobucar et al. 2013; Palermo and Thomson 2018; Roscoe and McNamara 2013; Stevenson 2016; Warschauer and Grimes 2008) and that students tend to be positive and less critical than teachers (Klobucar et al. 2013; Grimes and Warschauer 2010; Palermo and Thomson 2018; Roscoe et al. 2018). Moreover, younger students appear to be more trusting and more motivated than older students, an interpretation consistent with prior research showing that motivation declines as students progress in school (Nolen 2007).

Teachers in the present study showed evidence of holding nuanced attitudes towards AWE, which suggests that their judgements were free of halo or novelty effects. Teachers reported the least agreement—though still generally positive—with statements regarding the validity of the

automated scoring and the appropriateness of MI Write's automated feedback, but generally agreed that AWE is effective for a diverse range of learners including ELLs and students with disabilities, groups who historically struggle with writing (Graham et al. 2017; Salahu-Din et al. 2008).

Though our data do not reveal why teachers reported less agreement regarding the appropriateness of MI Write's automated feedback, anecdotal evidence suggests this may stem from differences in the way in which AWE and teachers provide feedback. AWE provides feedback in a consistent fashion, regardless of student effort or student characteristics, or at what point a student is within the curriculum (i.e., a student may receive feedback on skills not yet introduced by the teacher). This is quite different from the manner in which most teachers provide feedback and this may have led teachers to be less positive about MI Write's feedback (see Wilson et al. under review). Future research should more rigorously explore teachers' perceptions of automated feedback to understand the drivers and to identify ways in which AWE developers might improve automated feedback to support more successful instructional integration.

Interestingly, though teachers were the most skeptical about the congruence of MI Write's automated scoring with scores on Common Core-aligned rubrics and state test scores, the quantitative analyses showed that the PEG holistic score was the strongest predictor of state test writing performance, even stronger than many demographic factors and only weaker than prior state test performance. Prior research on screening with AWE has also shown automated scores to be valid predictors of state test performance (Wilson 2018; Wilson et al. 2016). Additional research should explore teachers' perceptions of automated scoring, how those perceptions are formed, and how they may later influence students' own perceptions of AWE systems (see Roscoe et al. 2018).

Associated Effects on Outcomes

The study extends prior research investigating AWE usage, particularly with the population of elementary students, and extends the limited research base related to gains in distal measures of writing performance, such as a state tests (see Shermis et al. 2004; Wilson and Roscoe 2019). Predominantly it was not AWE usage that was associated with writing outcomes; it was writing performance—as measured by MI Write's PEG automated essay scoring system—that most consistently predicted outcomes. Students' average first-draft PEG holistic score was a significant predictor of each outcome with effect sizes ranging from medium to large. AWE thus appears to index writing ability in ways that are meaningfully related to outcomes of interest and may hold promise for progress monitoring in natural settings, use as a benchmark writing assessment (Wilson et al. 2016), or periodic universal screening (Wilson 2018). This is an important finding as there is need for efficient, reliable, and comprehensive writing assessments to support teachers' instructional decision making (Graham et al. 2015).

The variable used as a proxy for writing practice (number of essays) predicted state test performance in Grades 3 and 5, with small effects in each model. The ability to improve one's writing in response to feedback (average gain score) predicted posttest writing quality and Grade 5 state test writing performance. However, the number of times students revised their essays (average drafts/essay) was not a statistically significant predictor of outcomes. This suggests that the number of revisions is not linearly correlated with improvements in writing quality. Indeed, prior research has shown that there is a saturation point for revision attempts in AWE at which point gains in writing quality are maximized (Wilson 2017; Wilson and Andrada 2016; Wilson et al. 2014). Instead, the average gain score, which measured how effectively students revised their writing, was a better predictor. While the average gain score is not a direct measure of AWE usage, it is an indirect measure of usage, reflecting the effort students invest to improve their performance

using automated feedback; indeed, the average gain score was moderately correlated with average drafts per essay in each grade. Nevertheless, even after accounting for covariates and classroom or school-level factors, the majority of student-level variance in outcomes was left unexplained by AWE usage and performance. Thus, it is important that future research continue to explore the direct and indirect effects of AWE usage on outcomes of interest.

Implications of Study Findings for the Design and Development of AWE Systems

A secondary goal of the present study was to gather validity evidence to evaluate the design decisions undergirding the architecture of MI Write. A key design decision was to balance scoring accuracy with feedback interpretability. Accuracy and interpretability can be at odds in artificial intelligence applications. For example, current deep learning approaches such as deep neural networks tend to be more accurate at predicting human ratings than simpler machine learning techniques. However, the complexity of the data transformations involved with such "black box" approaches introduces a challenge in relating model weights back to text features. This has the potential to maximize reliability (i.e., consistency) at the cost of validity, in particular construct validity. In contrast, interpretability is aided by defining a smaller set of instructionally-relevant features with clear connections to formative feedback. The architecture of MI Write attempts to balance accuracy and interpretability to foster user trust and encourage productive revising behaviors. Productive revising is further aided by PEG's feature-level saturation facet of the overall trait score; to realize recurrent improvements in performance across drafts, students must attend to errors and make improvements to their writing across levels of language and across traits.

Findings of analyses related to our third research question provide validity evidence in support of the theoretical and pedagogical decisions enacted in the architecture of MI Write. Specifically, though the average number of drafts per essay was positively and moderately

correlated with the average gain score, only the average gain score was predictive of outcomes. It was not the amount of revision, but how effectively a student revised that predicted performance on measures of writing proficiency. Thus, (a) because MI Write rewards revising strategies that are more comprehensive in nature (i.e., more effective revising strategies; MacArthur 2016), and (b) students with higher average gain scores performed better on measures of writing proficiency, irrespective of their average first-draft PEG score, this suggests that (c) MI Write is steering students to adopt skills and strategies that are beneficial on broader, distal measures of writing performance.

Yet, this was true only for two of the four outcomes: posttest PEG holistic score and fifth-grade Smarter Balanced writing performance. Neither average drafts per essay nor average gain score was predictive of state test performance in Grades 3 and 4. Additional supports may need to be embedded within MI Write, or provided by teachers, to scaffold effective revising and the acquisition of effective writing skills for third- and fourth-grade students. Scaffolding tools have shown promise for improving student writing (Rapp and Kauf 2018) and incorporating such tools within AWE systems may be a productive avenue of AWE development.

Study findings have several implications for the design of AWE systems. First, it is important that AWE systems enact a thoughtful balance of scoring accuracy and interpretability. Second, the design of scoring and feedback algorithms should be based around a pedagogically-sound theory of writing instruction (see also Roscoe et al. 2013). Third, scaffolds may be needed to support the revising behavior of AWE users (see also Roscoe et al. 2016), particularly younger students. Fourth, and finally, AWE developers should continue to examine associations between AWE usage and external measures in the context of naturalistic implementation. Such research

may provide, as it did in the current study, valuable evidence of the validity of underlying design decisions as well as insight into areas of system improvement.

Limitations and Future Directions

First, due to the nature of the RPP and the districtwide implementation of AWE, the study used a pretest-posttest design without a control group or random assignment. Evidence is thus correlational and cannot be interpreted as showing a causal relationship between AWE implementation and gains in writing outcomes. Further, in absence of a randomized control trial and A/B testing, we cannot say which specific features of MI Write were most effective, only that implementing the system as a whole was associated with the outcomes we observed. Our study also means that we cannot rule out the possibility that other instructional factors and not AWE usage per se contributed to the observed gains. However, even if the associated effects of AWE on outcomes reflect unmeasured instructional factors, it suggests that AWE was a part of an effective writing teacher's instructional approach. Nevertheless, to support claims about the impact of AWE implementation on outcomes future research should adopt more rigorous designs.

Second, we relied on log data recorded within MI Write and self-report teacher surveys to characterize MI Write implementation. Resource limitations prohibited conducting classroom observations to an extent that would allow for drawing representative conclusions. A productive avenue of future research would be to validate a classroom observation protocol for AWE-integrated writing instruction. Data from such a measure may help identify typologies of teachers' AWE implementation and those typologies may further explain variance in outcomes of interest.

Relatedly, though teachers were provided with standardized directions for administering the pretest and posttest writing prompts, we did not conduct checks on fidelity of assessment administration, electing instead to emphasize ecological validity given the focus of the study on

naturalistic implementation of AWE. Nevertheless, we cannot rule out the possibility that students' performance varied as a result of between-teacher differences in assessment administration.

Third, the 51% response rate to the spring teacher survey regarding AWE introduces the possibility that survey responders differed systematically from nonresponders. However, a series of ANOVA suggests that this may not be the case: There were no statistically significant differences in rates of classroom-level AWE usage between responders and nonresponders for number of assignments (F = 0.08, p = .772), class average drafts/essay (F = 0.04, p = .846), or class average first draft score (F = 0.45, p = .503). These non-significant differences mitigate the threat of response bias, but survey data should still be interpreted cautiously.

Fourth, we did not have access to teacher demographic data, and thus necessarily did not include relevant demographic variables as predictors in our models. However, we acknowledge that teacher characteristics, such as number of years teaching and degree of preparedness for teaching writing, may influence how teachers implement AWE. Future research should explore the interaction of teacher characteristics and AWE implementation. Such research may support efforts to improve targeted professional development and ongoing implementation support.

Fifth, we elected to examine the PEG holistic score as an outcome variable rather than the trait scores. Though our decision was grounded in practical (i.e., the holistic score is similar to other outcome measures) and psychometric reasons (i.e., high correlations among the trait and holistic scores), we acknowledge that our decision precluded us from determining whether improvements in overall writing quality were driven by gains in a subset of the traits versus all traits. Future research should explore the nature of gains in writing quality associated with use of AWE.

Sixth, our reliance on log data captured within MI Write meant that we were unable to account for the amount of writing that occurred outside the AWE system. Future AWE research should supplement log data with data regarding the number of assignments in the curriculum and the number of assignments teachers completed within and outside the AWE system. This would not only help to characterize the extent of AWE usage but would help to identify whether AWE usage is promoting writing practice over and above what is required by the curriculum.

Finally, in absence of an agreed-upon, highly specific definition of AWE we adopted a seminal definition of AWE that identifies automated scoring *and* automated feedback as central components of AWE (see Stevenson and Phakiti, 2014; Warschauer and Grimes, 2008). We studied a single AWE system, MI Write, that meets this definition. Our results are broadly aligned with prior research on other AWE systems that also conform to this definition. Yet, there are unique architectural differences in AWE systems, and given that we investigated MI Write only, results may not generalize to other AWE systems. Furthermore, our results would not generalize to programs, such as Grammarly, that solely provide error correction on low-level writing skills (e.g., spelling and grammar), despite such programs being categorized in recent years as AWE (e.g., Parra and Calero, 2019). As the field of AWE grows, it will be critical for researchers to attend to definitional issues surrounding AWE, including such issues as whether or not automated scoring is essential feature of AWE and whether any type of feedback, addressing any component of writing skill, qualifies a system as AWE. Having a clearer definition of AWE will improve the accuracy of claims made about the effectiveness, or lack thereof, of AWE, and will support practitioners seeking to adopt AWE to help improve the teaching and learning of writing.

References

Allison, P. D. (2009). *Fixed effects regression models.* Thousand Oaks, CA: SAGE.

Applebee, A. N., & Langer, J. A. (2009). What is happening in the teaching of writing? *English Journal, 98*(5), 18-28.

Archer, A. L., & Hughes, C. A. (2011). *Explicit instruction: Effective and efficient teaching.* New York, NY: Guilford.

Attali, Y. (2004, April). *Exploring the feedback and revision features of* Criterion. Paper presented at the National Council of Measurement in Education (NCME), San Diego, CA.

Bai, L., & Hu, G. (2017). In the face of fallible AWE feedback: How do students respond? *Educational Psychology, 37,* 67-81.

Bejar, I. I., Flor, M., Futagi, Y., & Ramineni, C. (2014). On the vulnerability of automated scoring to construct-irrelevant response strategies (CIRS): an illustration. *Assessing Writing, 22,* 48-59.

Bernoff, J. (2017, April 13). Bad writing costs businesses billions. *Daily Beast.* Retrieved from https://www.thedailybeast.com/bad-writing-costs-businesses-billions

Biber, D., Nekrasova, T., & Horn, B. (2011). *The effectiveness of feedback for L1-English and L2-writing development: A meta-analysis. TOEFL iBT™ Research Report.* Princeton, NJ: Educational Testing Service.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21,* 5-31.

Britt, M. A., Wiemer-Hastings, P., Larson, A. A., & Perfetti, C. A. (2004). Using intelligent feedback to improve sourcing and integration in students' essays. *International Journal of Artificial Intelligence in Education, 14,* 359-374.

Bruning, R., Dempsey, M., Kauffman, D. F., McKim, C., & Zumbrunn, S. (2013). Examining dimensions of self-efficacy for writing. *Journal of Educational Psychology, 105*, 25-38.

Bruning, R. H., & Kauffman, D. F. (2016). Self-efficacy beliefs and motivation in writing development. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of Writing Research* (pp. 160-173). New York, NY: Guilford.

Bunch, M. B., Vaughn, D., & Miel, S. (2016). Automated scoring in assessment systems. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on technology tools for real-world skill development* (pp. 611-626). Hershey, PA: IGI Global.

Caccamise, D., Franzke, M., Eckhoff, A., Kintsch, E., & Kintsch, W. (2007). Guided practice in technology-based summary writing. In D. S. McNamara, *Reading comprehension strategies: Theories, interventions, and technologies* (pp. 375-396). Mahwah, NJ: Erlbaum.

Chapelle, C. A., Cotos, E., & Lee, J. (2015). Validity arguments for diagnostic assessment using automated writing evaluation. *Language Testing, 33*, 385-405.

Chen, C. E., & Cheng, W. E. (2008). Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes. *Language Learning & Technology, 12*(2), 94-112.

Clare, L., Valdés, R., & Patthey-Chavez, G. G. (2000). *Learning to write in urban elementary and middle schools: An investigation of teachers' written feedback on student compositions* (Center for the Study of Evaluation Technical Report No. 526). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Cohen J. (1988). *Statistical Power Analysis for the Behavioral Sciences.* New York, NY: Routledge Academic.

Conference on College Composition and Communication. (2014). *Writing assessment: A position statement.* Retrieved from https://ncte.org/statement/writingassessment/.

Dujinhower, H., Prins, F. J., & Stokking, K. M. (2012). Feedback providing improvement strategies and reflection on feedback use: Effects on students' writing motivation, process, and performance. *Learning and Instruction, 22,* 171-184.

Ericcson, P. F., & Haswell, R. J. (2006). *In Machine scoring of student essays: Truth and consequences.* Logan, UT: Utah State University Press.

Fitzgerald, J. (1987). Research on revision in writing. *Review of Educational Research, 57,* 481-506.

Foltz, P. W., Lochbaum, K. E. & Rosenstein, M. B. (2011, April). *Analysis of student ELA writing performance for a large scale implementation of formative assessment.* Paper presented at the Annual Meeting of the National Council for Measurement in Education, New Orleans, LA.

Foltz, P. W., Streeter, L. A., Lochbaum, K. E., & Landauer, T. K. (2013). Implementation and applications of the Intelligent Essay Assessor. In M. Shermis, & J. Burstein (Eds.), *Handbook of Automated Essay Evaluation* (pp. 68-88). New York, NY: Routledge.

Franzke, M., Kintsch, E., Caccamise, D., Johnson, N., and Dooley, S. (2005). Summary Street®: Computer support for comprehension and writing. *Journal of Educational Computing Research, 33,* 53–80.

Gansle, K. A., VanDerHeyden, A. M., Noell, G. H., Resetar, J. L., & Williams, K. L. (2006). The technical adequacy of curriculum-based and rating-based measures of written expression of elementary school students. *School Psychology Review, 35*, 435–450.

Gilbert, J., & Graham, S. (2010). Teaching writing to elementary students in grades 4-6: A national survey. *Elementary School Journal, 110*, 494-518.

Graham, S. (2018). A revised writer(s)-within-community model of writing. *Educational Psychologist, 53*, 258-279.

Graham, S., Berninger, V., & Abbott, R. (2012). Are attitudes toward writing and reading separable constructs? A study with primary grade children. *Reading & Writing Quarterly, 28*, 51-69.

Graham, S., Berninger, V., & Fan, W. (2007). The structural relationship between writing attitude and writing achievement in first and third grade students. *Contemporary Educational Psychology*, *32*(3), 516–536.

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.

Graham, S., Collins, A. A., Rigby-Wells, H. (2017). Writing characteristics of students with learning disabilities and typically achieving peers: A meta-analysis. *Exceptional Children, 83*, 199-218.

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

Graham, S., & Perin, D. (2007). *Writing next: Effective strategies to improve writing of adolescents in middle and high schools – A report to Carnegie Corporation of New York*. Washington, DC: Alliance for Excellent Education.

Grimes, D., & Warschauer, M. (2010). Utility in a fallible tool: A multi-site case study of automated writing evaluation. *Journal of Technology, Learning, and Assessment, 8*(6), 4–43.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81–112.

Hayes, J. R. (2012). Modeling and remodeling writing. *Written Communication, 29*(3), 369–388.

Herrington, A., & Moran, C. (2001). What happens when machines read our students' writing? *College English, 63*(4), 480-499.

Higgins, D., & Heilman, M. (2014). Managing what we can measure: quantifying the susceptibility of automated scoring systems to gaming behavior. *Educational Measurement Issues and Practice, 33*(3), 36-46.

Hornick, K. 1991. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, *4*, 359–366

Kellogg, R. T., & Whiteford, A. P. (2009). Training advanced writing skills: The case for deliberative practice. *Educational Psychologist, 44*, 250-266.

Kellogg, R. T., Whiteford, A. P., & Quinlan, T. (2010). Does automated feedback help students learn to write? *Journal of Educational Computing Research, 42*, 173-196.

Kiuhara, S. A., Graham, S., & Hawken, L. S. (2009). Teaching writing to high school students: A national survey. *Journal of Educational Psychology, 101*, 136-160.

Klobucar, A., Elliot, N., Deess, P., Rudniy, O., & Joshi, K. (2013). Automated scoring in context: Rapid assessment for placed students. *Assessing Writing, 18,* 62-84.

Lee, V. (2000). Using hierarchical linear modeling to study social Contexts: The case of school effects, *Educational Psychologist, 35*, 125-141.

Lorah, J. (2018). Effect size measures for multilevel models: Definition, interpretation, and TIMSS example. *Large-Scale Assessments in Education*, *6*(8), 1-11.

MacArthur, C. A. (2016). Instruction in evaluation and revision. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research, 2nd Ed.* (pp. 272-287). New York, NY: Guilford.

Matsumara, L. C., Patthey-Chavez, G. G., Patthey-Chavez, Valdes, M. R., & Garnier, H. (2002). Teacher feedback, writing assignment quality, and third-grade students' revision in lower- and higher-achieving urban schools. *Elementary School Journal, 103,* 3-25.

Mayfield, E., & Butler, S. (2018). *Districtwide implementations outperform isolated use of automated feedback in high school writing.* Paper presented at the International Conference of the Learning Sciences, London, United Kingdom. Retrieved from http://ceur-ws.org/Vol-2128/industrial4.pdf

Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record, 108*, 1017-1054.

Moore, N. S., & MacArthur, C. A. (2016). Student use of automated essay evaluation technology during revision. *Journal of Writing Research, 8*, 149-175.

National Center for Education Statistics (2012).*The Nation's Report Card: Writing 2011*(NCES 2012–470). Institute of Education Sciences, U.S. Department of Education, Washington, D.C.

National Commission on Writing for America's Families, Schools, and Colleges. (2003). *The neglected "R": The need for a writing revolution.* Iowa City, IA: The College Board

National Commission on Writing for America's Families, Schools, and Colleges. (2004). *Writing: A ticket to work…or a ticket out. A survey of business leaders.* Iowa City, IA: The College Board.

National Commission on Writing for America's Families, Schools, and Colleges. (2005). *Writing: A powerful message from state government.* Iowa City, IA: The College Board.

National Council of Teachers of English. (2013). *NCTE position statement on machine scoring*. Retrieved from http://www.ncte.org/positions/statements/machine_scoring

Newman, D., Jaciw, A. P., & Lazarev, V. (2018). *Guidelines for conducting and reporting EdTech impact research in U.S. K-12 schools.* Empirical Education. Retrieved from https://www.empiricaleducation.com/pdfs/guidelines.pdf

Nguyen, H., Xiong, W., & Litman, D. (2017). Iterative design and classroom evaluation of automated feedback for improving peer feedback localization. *International Journal of Artificial Intelligence in Education, 27*, 582-622.

Northwest Regional Educational Laboratory. (2004). *An introduction to the 6+1 trait writing assessment model*. Portland, OR: Author.

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

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.

Parra, G. L., & Calero, S. X. (2019). Automated writing evaluation tools in the improvement of the writing skill. *International Journal of Instruction, 12*(2), 209-226. https://doi.org/10.29333/iji.2019.12214a

Parr, J. M., & Timperley, H. S. (2010). Feedback to writing, assessment for teaching and learning and student progress. *Assessing Writing, 15,* 68-85.

Perelman, L. (2014). When the "state of the art" is counting words. *Assessing Writing, 21,* 104-111. http://dx.doi.org/10.1016/j.asw.2014.05.001.

Persky, H. R., Daane, M. C., & Jin, Y. (2002). *The Nation's Report Card: Writing 2002.* (NCES 2003-529). National Center for Education Statistics, Institute of Education Sciences. U. S. Department for Education, Washington, D. C.

Rapp, C., & Kauf, P. (2018). Scaling academic writing instruction: Evaluation of a scaffolding tool (Thesis Writer). *International Journal of Artificial Intelligence in Education, 28,* 590-615.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

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/10.1177/1541931218621471

Roscoe, R. D., Jacovina, M. E., Allen, L. K., Johnson, A. C., & McNamara, D. S. (2016). Towards revision-sensitive feedback in automated writing evaluation. *Proceedings of the 9th International Conference on Educational Data Mining,* 628-629.

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*, 1010-1025.

Roscoe, R. D., Varner, L. K., Crossley, S. A., McNamara, D. S. (2013). Developing pedagogically-guided algorithms for intelligent writing feedback. *International Journal of Learning Technology, 8*, 362-381.

Roscoe, R. D., Wilson, J., Johnson, A. C., & Mayra, C. R. (2017). Presentation, expectations, and experience: Sources of student perceptions of automated writing evaluation. *Computers in Human Behavior, 70*, 207-221.

Salahu-Din, D., Persky, H., and Miller, J. (2008). *The Nation's Report Card: Writing 2007* (NCES 2008–468). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, Washington, D.C.

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.

Shermis, M. D., Burstein, J. C., & Bliss, L. (2004, April). *The impact of automated essay scoring on high stakes writing assessments.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

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

Shermis, M. D., Koch, C. M., Page, E. B., Keith, T. Z., & Harrington, S. (2002). Trait ratings for automated essay grading. *Educational and Psychological Measurement, 62*, 5-18.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78,* 153–189.

Smarter Balanced Assessment Consortium. (2018). Smarter Balanced Assessment Consortium: 2017-18 summative technical report. Retrieved from https://www.smarterbalanced.org/wp-content/uploads/2019/08/2017-18-Summative-Assessment-Technical-Report.pdf

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and* advanced *multilevel modeling.* Thousand Oaks: Sage Publishing.

Stevenson, M. (2016). A critical interpretative synthesis: The integration of Automated Writing Evaluation into classroom writing instruction. *Computers and Composition, 42*, 1-16.

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

Tansomboon, C., Gerard, L. F., Vitale, J. M., & Linn, M. C. (2017). Designing automated guidance to promote productive revision of science explanations. *International Journal of Artificial Intelligence in Education, 27*, 729-757.

Troia, G. A., Harbaugh, A. G., Shankland, R. K., Wolbers, K. A., & Lawrence, A. M. (2013). Relationships between writing motivation, writing activity, and writing performance: Effects of grade, sex, and ability. *Reading and Writing, 26*, 17-44.

Troia, G. A., Olinghouse, N. G., Zhang, M., Wilson, J., Stewart, K. A., Mo, Y., & Hawkins, L. (2018). Content and alignment of state writing standards and assessments as predictors of student writing achievement: An analysis of 2007 National Assessment of Educational Progress data. *Reading and Writing, 31*, 835-864.

Warschauer, M., & Grimes, D. (2008). Automated writing assessment in the classroom. *Pedagogies, 3,* 22–36.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*, 2–13.

Wilson, J. (2017). Associated effects of automated essay evaluation software on growth in writing quality for students with and without disabilities. *Reading and Writing, 30,* 691-718.

Wilson, J. (2018). Universal screening with automated essay scoring: Evaluating classification accuracy in Grades 3 and 4. *Journal of School Psychology, 68,* 19-37.

Wilson, J., Ahrendt, C., Fudge, E., Raiche, A., Beard, G., & MacArthur, C. A. (Under Review). *Elementary teachers' perceptions of automated feedback and automated scoring: Transforming the teaching and learning of writing using automated writing evaluation.*

Wilson, J., & Andrada, G. N. (2016). Using automated feedback to improve writing quality: Opportunities and challenges. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on technology tools for real-world skill development* (pp.678-703). Hershey, PA: IGI Global.

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.

Wilson, J., Olinghouse N. G., & Andrada, G. N. (2014). Does automated feedback improve writing quality? *Learning Disabilities: A Contemporary Journal, 12*, 93-118.

Wilson, J., Olinghouse, N. G., McCoach, D. B., Andrada, G. N., & Santangelo, T. (2016). Comparing the accuracy of different scoring methods for identifying sixth graders at risk of failing a state writing assessment. *Assessing Writing*, *27*, 11-23.

Wilson, J., & Roscoe, R. D. (2019). Automated writing evaluation and feedback: Multiple metrics of efficacy. *Journal of Educational Computing Research.* Advance online publication. http://dx.doi.org/10.1037/edu0000311

Zellermayer, M., Salomon, G., Globerson, T., & Givon, H. (1991). Enhancing writing-related metacognition through a computerized writing partner. *American Educational Research Journal, 28,* 373-391.

Zhang, H., Magooda, A., Litman, D., Correnti, R., Wang, E., Matsumara, L. C.,…Quintanta, R. (2019). eRevise: Using natural language processing to provide formative feedback on text evidence usage in student writing. *Proceedings of the AAAI Conference on Innovative Applications of Artificial Intelligence, 33*, 9619-9625.

Table 1

Demographics of Student Sample

Variable	Number	Percentage
Grade		
3	668	34.5
4	747	38.6
5	520	26.9
Gender		
Male	955	49.4
Female	980	50.6
Race		
African American	300	15.5
Asian	130	6.7
Hispanic/Latino	606	31.3
White	1490	77.0
Others	4	0.2
SPED	207	10.7
ELL	437	22.6

Note. N = 1935. Racial categories were not mutually exclusive; therefore, the percentages total to more than 100 percent.

Table 2

Descriptive Statistics for MI Write Usage, Self-Efficacy, and Outcome Variables

	Overall *M (SD)*	Grade 3 *M (SD)*	Grade 4 *M (SD)*	Grade 5 *M (SD)*
Class Number of Assignments	4.44 (3.16)	3.96 (2.26)	4.82 (3.93)	4.56 (2.99)
Class Average Drafts/Essay	3.70 (2.61)	3.27 (2.31)	3.48 (2.32)	4.62 (3.21)
Class Average Score across Prompts	16.84 (4.50)	15.32 (4.04)	17.44 (4.75)	18.10 (4.27)
Student Total Number of Essays	6.84 (5.39)	5.23 (3.28)	8.14 (6.80)	7.03 (4.73)
Student Average Drafts/Essay	4.56 (3.36)	4.02 (3.18)	4.38 (3.03)	5.49 (3.80)
Student Average First Draft Score	15.28 (3.82)	14.50 (3.47)	15.39 (3.93)	16.12 (3.90)
Student Average Gain Score	1.85 (1.75)	1.47 (1.55)	1.90 (1.76)	2.27 (1.87)
Student Lesson Minutes (1−12mins)	0.16 (0.37)	0.18 (0.38)	0.13 (0.34)	0.19 (0.40)
Student Lesson Minutes (≥ 13mins)	0.24 (0.43)	0.23 (0.42)	0.28 (0.45)	0.21 (0.41)
Average Self-Efficacy	4.02 (0.60)	3.98 (0.64)	4.02 (0.59)	4.08 (0.55)
Pretest PEG Holistic Score	15.31 (5.11)	13.31 (4.57)	16.57 (5.28)	16.09 (4.74)
Posttest PEG Holistic Score	18.24 (4.75)	17.14 (4.68)	18.89 (4.95)	18.71 (4.30)
2017 Smarter Balanced Writing Score[a]	2456.51 (92.31)	-	2441.54 (90.93)	2479.52 (88.80)
2018 Smarter Balanced Writing Score[a]	2481.33 (110.69)	2436.91 (98.38)	2481.46 (110.10)	2538.21 (100.05)

Note. $N_{student}$*:* Overall = 1935; Grade 3 = 668; Grade 4 = 729; Grade = 520. $N_{teacher}$: Overall = 135; Grade 3 = 49; Grade 4 = 51; Grade 5 = 35. [a]Smarter Balanced writing scale score (range = 2000–3000).

Table 3

Teachers' Attitudes toward AWE

Item	*M (SD)*	*Mdn*	Mode
Ease of Use and Acceptability			
1. MI Write is easy for me to use.	3.75 (0.85)	4	4
2. MI Write is easy for my students to use.	3.65 (1.00)	4	4
3. MI Write scores correlate with Common Core rubrics.	3.29 (0.94)	3	4
4. MI Write scores are accurate predictors of Smarter Balanced.	3.04 (0.93)	3	3
5. MI Write's feedback is appropriate.	3.28 (1.07)	4	4
6. MI Write helps me differentiate instruction.	3.33 (0.97)	4	4
7. I would like to continue using MI Write.	4.00 (1.03)	4	4
Effects on student learning			
8. MI Write helps students improve their writing.	3.62 (1.00)	4	4
9. MI Write increases writing motivation.	3.81 (1.05)	4	4
10. ELLs benefit from MI Write.	3.46 (0.96)	4	4
11. Students with disabilities benefit from MI Write.	3.48 (0.98)	4	4
12. Students write and revise more with MI Write.	3.93 (1.06)	4	4
13. Students receive more feedback with MI Write.	3.88 (0.93)	4	4
14. Students are able to apply the feedback they receive from MI Write.	3.62 (1.07)	4	4
Effects on Instruction			
15. MI Write helps me address my students' needs.	3.70 (0.79)	4	4
16. I assign more writing when I use MI Write.	3.33 (1.11)	4	4
17. Using MI Write helps me teach more and grade less.	3.59 (1.06)	4	4
18. Using MI Write saves me time.	3.72 (1.08)	4	4
19. Using MI Write lets me focus on higher concerns of writing instead of mechanics.	3.78 (0.92)	4	4
20. I am effectively using MI Write to teach writing.	3.33 (1.04)	4	4

Note. $N = 69$. Response range = 1−5. 1 = Definitely not; 2 = Probably not; 3 = I don't know; 4 = Probably yes; 5 = Definitely yes.

Table 4

Students' Attitudes toward AWE

Item	Overall			Grade 3	Grade 4	Grade 5
	M (SD)	*Mdn*	Mode	*M (SD)*	*M (SD)*	*M (SD)*
1. MI Write helps me become a better writer.	3.72 (1.24)	4	5	3.83 (1.22)	3.67 (1.25)	3.64 (1.22)
2. MI Write helps me learn to revise my writing.	3.86 (1.23)	4	5	3.85 (1.21)	3.86 (1.26)	3.88 (1.22)
3. I believe the scores that MI Write gives	3.66 (1.34)	4	5	3.80 (1.29)	3.57 (1.40)	3.62 (1.28)
4. I fix my writing more when I use MI Write.	3.81 (1.30)	4	5	3.75 (1.32)	3.84 (1.29)	3.85 (1.29)
5. MI Write helps me feel more motivated to write.	3.32 (1.39)	4	5	3.40 (1.37)	3.31 (1.43)	3.21 (1.36)
6. I understand the feedback from MI Write.	3.61 (1.29)	4	5	3.62 (1.27)	3.57 (1.35)	3.67 (1.24)
7. MI Write helps me learn what parts of my writing I need to work on.	3.92 (1.25)	4	5	3.94 (1.27)	3.94 (1.24)	3.86 (1.24)
8. I understand how to use MI Write.	4.38 (1.11)	5	5	4.30 (1.18)	4.42 (1.06)	4.44 (1.04)
9. I want to use MI Write next year.	3.41 (1.51)	4	5	3.63 (1.45)	3.35 (1.55)	3.23 (1.50)

Note. Response range = 1–5. 1 = Definitely not; 2 = Probably not; 3 = I don't know; 4 = Probably yes; 5 = Definitely yes.

Table 5

HLM Results Predicting Posttest PEG Holistic Score

	Model 1 – Unconditional with School Dummy Codes		Model 2 – Student Level Predictors		Model 3 – Student Level and Teacher-Level Predictors		
Fixed Effects	***Coefficient B* (S.E.)**	***t***	***Coefficient B* (S.E.)**	***t***	***Coefficient* B (S.E.)**	***t***	***Effect Size (β)***
Intercept	17.96*** (0.22)	81.75	17.95*** (0.22)	81.19	17.87*** (0.20)	87.26	
School 1	-0.95 (1.14)	-0.84	-0.90 (1.14)	-0.79	-0.23 (1.08)	-0.21	-0.01
School 2	2.01$^{\sim}$ (1.02)	1.04	2.10* (1.05)	2.00	2.50* (0.98)	2.56	0.16*
School 3	0.08 (1.04)	1.93	0.21 (1.04)	0.20	-0.24 (0.98)	-0.24	-0.02
School 4	-0.50 (1.08)	-0.47	-0.39 (1.08)	-0.36	-0.42 (1.01)	-0.42	-0.02
School 5	2.00$^{\sim}$ (1.08)	1.85	2.08 (1.09)	1.91	1.61 (1.02)	1.57	0.09
School 6	-1.34 (1.30)	-1.04	-1.28 (1.29)	-0.99	-1.12 (1.22)	-0.92	-0.05
School 7	-1.83 (1.37)	-1.33	-1.72 (1.38)	-1.25	-0.75 (1.29)	-0.58	0.03
School 8	0.22 (1.17)	0.19	0.32 (1.18)	0.27	0.05 (1.11)	0.05	0.00
School 9	2.30* (1.15)	1.99	2.46* (1.16)	2.12	1.91 (1.10)	1.73	0.10
School 10	-1.07 (1.18)	-0.91	-1.00 (1.18)	-0.84	-1.06 (1.10)	-0.97	-0.05
School 11	2.21 (1.14)	1.95	2.28* (1.14)	2.00	1.89 (1.11)	1.70	0.11
School 12	-2.20 (1.12)	-1.96	-2.12 (1.12)	-1.89	-1.28 (1.06)	-1.21	-0.08
School 13	-3.28 (2.06)	-1.59	-3.27 (1.96)	-1.66	-2.50 (1.89)	-1.32	-0.08
	df = 117		*df* = 117			*df* = 114	
Level-1 Predictors							
Grade			0.45 (2.00)	0.23	0.45 (2.00)	0.23	0.07
Gender			-0.90*** (0.16)	-5.50	-0.90*** (0.16)	-5.51	-0.09***
Hispanic			0.10 (0.25)	0.39	0.10 (0.25)	0.39	0.01
African American			-0.38 (0.25)	-1.52	-0.38 (0.25)	-1.52	-0.03
Asian			1.50*** (0.36)	4.19	1.50*** (0.36)	4.19	0.08***
Special Education			-0.99*** (0.27)	-3.62	-0.99*** (0.27)	-3.62	-0.06***
English Language Learner			-0.07 (0.27)	-0.27	-0.07 (0.27)	-0.27	-0.01
Student Total Number of Essays			0.01 (0.02)	0.48	0.01 (0.02)	0.48	0.01
Student Average Drafts/Essay			-0.02 (0.04)	-0.59	-0.02 (0.04)	-0.59	-0.01
Student Average First Draft Score			0.32*** (0.03)	11.72	0.32*** (0.03)	11.73	0.26***
Student Average Gain Score			0.34*** (0.07)	4.82	0.34*** (0.07)	4.82	0.12***
Total Lesson Minutes (1–12)			-0.26 (0.25)	-1.06	-0.26 (0.25)	-1.06	-0.02
Total Lesson Minutes (≥13)			0.01 (0.25)	-1.06	0.01 (0.25)	-1.06	0.00
Pretest PEG Holistic Score			0.26*** (0.02)	11.46	0.26*** (0.02)	11.46	0.28***
Average Self-Efficacy Score			0.60*** (0.15)	4.07	0.60*** (0.15)	4.08	0.08***
			df = 1725			*df* = 1725	
Level-2 Predictors							
Teachers' Number of Assignments					0.05 (0.08)	0.54	0.03
Class Average Drafts/Essay					0.08 (0.10)	0.74	0.04
Class Average Score across Prompts					0.23*** (0.06)	3.72	0.22***
						df = 114	

Variance Components						
r_{ij}: Level-1 (students)	16.72		11.32		11.31	
u_{0j}: Level-2 (teachers)	4.70		5.23		4.32	
Deviance	10742.40	$df = 2$	10084.82	$df = 2$	10070.17	$df = 2$
AIC	10746.40		10088.82		10074.17	
BIC	10757.47		10099.89		10085.24	
SBIC	10751.11		10093.53		10078.88	
Percent Variance Explained			Level-1 (student)		32.33%	
			Level-2 (teachers)		8.20%	

Note. $^{*}p < 0.05$. $^{**}p < 0.01$. $^{***}p < 0.001$.

Table 6

HLM Results Predicting 2018 Grade 3 State Test Writing Performance

	Model 1 – Unconditional with School Dummy Codes		**Model 2 – Student Level Predictors**		**Model 3 – Student Level and Teacher-Level Predictors**		
Fixed Effects	***Coefficient B* (S.E.)**	***t***	***Coefficient B* (S.E.)**	***t***	***Coefficient* B (S.E.)**	***t***	***Effect Size (ß)***
Intercept	2427.68*** (13.04)	186.24	2421.42*** (13.24)	182.88	2416.39*** (7.13)	339.11	
	df = 13		*df* = 13		*df* = 8		
Level-1 Predictors							
Gender			-10.69 (6.06)	-1.76	-10.71 (6.06)	-1.77	-0.05
Hispanic			2.92 (9.65)	0.30	2.94 (9.65)	0.31	0.01
African American			-35.84*** (9.45)	-3.79	-35.85*** (9.44)	-3.80	-0.13***
Asian			52.60*** (11.94)	4.41	52.63*** (11.93)	4.41	0.15***
Special Education			-49.63*** (10.15)	-4.89	-49.60*** (10.15)	-4.89	-0.15***
English Language Learner			-54.20*** (9.89)	-5.48	-54.22*** (9.88)	-5.49	-0.24***
Student Total Number of Essays			2.29* (1.04)	2.20	2.29* (1.04)	2.20	0.08*
Student Average Drafts/Essay			-1.58 (1.30)	-1.21	-1.60 (1.30)	-1.23	-0.05
Student Average First Draft Score			7.20*** (0.94)	7.63	7.20*** (0.94)	7.63	0.25***
Student Average Gain Score			4.13 (2.65)	1.56	4.22 (2.64)	1.60	0.07
Total Lesson Minutes (1–12)			-1.47 (8.56)	-0.17	-1.48 (8.55)	-0.17	-0.01
Total Lesson Minutes (≥13)			-0.11 (7.85)	-0.01	-0.11 (7.85)	-0.01	0.00
Average Self-Efficacy Score			18.54*** (5.11)	3.63	18.53*** (5.11)	3.63	0.12***
			df = 641		*df* = 641		
Level-2 Predictors							
Aggregated Teachers' Number of Assignments					7.04 (8.15)	0.86	0.12
Aggregated Class Average Drafts/Essay					0.32 (8.09)	0.04	0.00
Aggregated Class Average Score across Prompts					-10.31 (6.42)	-1.61	-0.21
School Free Reduced-Price Lunch Percentage					-2.26*** (0.42)	-5.44	-0.57***
School Adequate ELA Growth Percentage					-1.01 (1.21)	-0.83	-0.08
					df = 8		
Variance Components							
r: Level-1 (students)	7746.23		5782.29		5778.71		
u_0: Level-2 (schools)	2030.45		1978.82		326.88		
Deviance	7901.15	*df* = 2	7634.60	*df* = 2	7592.06	*df* = 2	
AIC	7905.15		7638.60		7596.06		

BIC	7914.16	7647.61	7605.07
SBIC	7907.81	7641.26	7598.72
Percent Variance Explained	Level-1 (student)		25.40%
	Level-2 (schools)		83.90%

Note. $^{*}p < 0.05$. $^{**}p < 0.01$. $^{***}p < 0.001$.

Table 7

HLM Results Predicting 2018 Grade 4 State Test Writing Performance

	Model 1 – Unconditional with School Dummy Codes		**Model 2 – Student Level Predictors**		**Model 3 – Student Level and Teacher-Level Predictors**		
Fixed Effects	***Coefficient B* (S.E.)**	***t***	***Coefficient B* (S.E.)**	***t***	***Coefficient* B (S.E.)**	***t***	***Effect Size (β)***
Intercept	2470.98*** (12.37)	199.78	2468.78*** (13.30)	185.60	2466.47*** (7.69)	320.57	
	df = 13		*df* = 13		*df* = 8		
Level-1 Predictors							
Gender			-8.51 (5.72)	-1.49	-8.51 (5.72)	-1.49	-0.04
Hispanic			3.61 (8.74)	0.41	3.61 (8.73)	0.41	0.02
African American			-10.39 (8.64)	-1.20	-10.39 (8.63)	-1.20	-0.03
Asian			21.38 (13.72)	1.56	21.38 (13.71)	1.56	0.04
Special Education			-32.16*** (9.18)	-3.50	-32.16*** (9.17)	-3.51	-0.10***
English Language Learner			-19.25* (9.46)	-2.04	-19.25* (9.45)	-2.04	-0.08*
Student Total Number of Essays			0.23 (0.50)	0.46	0.23 (0.50)	0.46	0.01
Student Average Drafts/Essay			-2.12 (1.37)	-1.55	-2.12 (1.37)	-1.55	-0.06
Student Average First Draft Score			3.08*** (0.81)	3.80	3.08*** (0.81)	3.81	0.11***
Student Average Gain Score			1.93 (2.34)	0.83	1.93 (2.34)	0.83	0.03
Total Lesson Minutes (1–12)			1.30 (8.76)	0.15	1.30 (8.76)	0.15	0.00
Total Lesson Minutes (≥13)			2.20 (7.45)	0.30	2.20 (7.44)	0.30	0.01
2017 Writing Score			0.58*** (0.03)	16.64	0.58*** (0.03)	16.66	0.55***
Average Self-Efficacy Score			12.33* (5.00)	2.47	12.33* (5.00)	2.47	0.07*
			df = 719		*df* = 719		
Level-2 Predictors							
Aggregated Teachers' Number of Assignments					-6.30 (10.46)	-0.60	-0.09
Aggregated Class Average Drafts/Essay					2.98 (10.84)	0.28	0.04
Aggregated Class Average Score across Prompts					-1.29 (8.36)	-0.16	-0.02
School Free Reduced-Price Lunch Percentage					-1.71* (0.53)	-3.22	-0.39***
Academic Progress Growth in ELA Percentage					1.99 (1.53)	-1.30	0.14
					df = 8		
Variance Components							
r: Level-1 (students)	10282.69		5544.78		5533.45		
u_0: Level-2 (schools)	1982.65		2247.27		643.23		
Deviance	9044.90	*df* = 2	8523.15	*df* = 2	8481.13	*df* = 2	

AIC	9048.90	8527.15	8485.13
BIC	9058.13	8536.38	8494.36
SBIC	9051.78	8530.03	8488.01
Percent Variance Explained	Level-1 (student)		46.19%
	Level-2 (schools)		67.56%

Note. $^{*}p < 0.05$. $^{**}p < 0.01$. $^{***}p < 0.001$.

Table 8

HLM Results Predicting 2018 Grade 5 State Test Writing Performance

	Model 1 – Unconditional with School Dummy Codes		**Model 2 – Student Level Predictors**		**Model 3 – Student Level and Teacher-Level Predictors**		
Fixed Effects	***Coefficient B* (S.E.)**	***t***	***Coefficient B* (S.E.)**	***t***	***Coefficient* B (S.E.)**	***t***	***Effect Size (β)***
Intercept	2532.13*** (13.87)	182.60	2532.23*** (14.48)	174.83	2530.18*** (5.34)	473.85	
	df = 12		*df* = 12		*df* = 7		
Level-1 Predictors							
Gender			-1.43 (5.81)	-0.25	-1.43 (5.80)	-0.25	-0.01
Hispanic			0.59 (7.68)	0.08	0.59 (7.66)	0.08	0.00
African American			-9.35 (9.10)	-1.03	-9.35 (9.08)	-1.03	-0.03
Asian			36.82** (12.29)	3.00	36.82** (12.27)	3.00	0.10**
Special Education			-28.80** (10.46)	-2.75	-28.80** (10.44)	-2.76	-0.08**
English Language Learner			-25.82** (9.39)	-2.75	-25.82** (9.38)	-2.75	-0.09**
Student Total Number of Essays			1.52* (0.65)	2.34	1.52* (0.65)	2.35	0.07*
Student Average Drafts/Essay			-1.30 (0.98)	-1.33	-1.30 (0.97)	-1.34	-0.05
Student Average First Draft Score			4.80*** (0.79)	6.04	4.80*** (0.79)	6.05	0.19***
Student Average Gain Score			4.07* (2.01)	2.02	4.07* (2.01)	2.03	0.08*
Total Lesson Minutes (1–12)			4.70 (7.87)	0.60	4.70 (7.85)	0.60	0.02
Total Lesson Minutes (≥13)			14.55 (7.46)	1.95	14.55 (7.45)	1.95	0.06
2017 Writing Score			0.50*** (0.03)	15.50	0.50*** (0.03)	15.53	0.51***
Average Self-Efficacy Score			15.04** (5.52)	2.73	15.04** (5.51)	2.73	0.08**
			df = 493		*df* = 493		
Level-2 Predictors							
Aggregated Teachers' Number of Assignments					16.71 (7.24)	2.31	0.12
Aggregated Class Average Drafts/Essay					-16.64 (7.06)	-2.36	0.00
Aggregated Class Average Score across Prompts					-6.55 (5.36)	-1.22	-0.21
School Free Reduced-Price Lunch Percentage					-2. 46*** (0.37)	-7.17	-0.57***
Academic Progress Growth in ELA Percentage					-1.47 (1.03)	-1.43	-0.08
					df = 7		
Variance Components							
r_0 Level-1 (students)	8450.95		3915.90		3903.69		

u_0: Level-2 (schools)	2112.81		2518.77		211.15	
Deviance	6196.18	*df* = 2	5737.69	*df* = 2	5693.45	*df* = 2
AIC	6200.18		5741.69		5697.45	
BIC	6208.69		5750.20		5705.96	
SBIC	6202.34		5743.85		5699.61	
Percent Variance Explained		Level-1 (student)			53.81%	
		Level-2 (schools)			90.01%	

Note. $^{*}p < 0.05$. $^{**}p < 0.01$. $^{***}p < 0.001$.

Table 9

Summary of Standardized Coefficients for AWE-usage Variables Predicting Proximal and Distal Outcomes

	Posttest Writing Quality (PEG Holistic Score)	2018 Grade 3 Smarter Balanced Writing Scale Score	2018 Grade 4 Smarter Balanced Writing Scale Score	2018 Grade 5 Smarter Balanced Writing Scale Score
Student-level Predictors				
Student Total Number of Essays	0.01	0.08*	0.01	0.07*
Student Average Drafts/Essay	−0.01	−0.05	−0.06	−0.05
Student Average First-Draft Score	0.26***	0.25***	0.11***	0.19***
Student Average Gain Score	0.12***	0.07	0.03	0.08*
Student Total Lesson Minutes (1–12)	−0.02	−0.01	0.00	0.02
Student Total Lesson Minutes (≥13)	0.00	0.00	0.01	0.06
Variance Explained (%)	32.33%	25.40%	46.19%	53.81%
Teacher-level Predictors[a]				
Class Total Essays Assigned	0.03	0.12	−0.09	0.12
Class Average drafts/essay	0.04	0.00	0.04	0.00
Class Average Score across Prompts	0.22***	−0.21	−0.02	−0.21
Variance Explained (%)	8.20%	83.90%	67.56%	90.01%

Note. According to Cohen's (1988) interpretation of standardized coefficients, we identify .05 as a small effect, .10 as a medium effect, and .25 as a large effect. Variance explained (%) refers to the full conditional model with additional predictors besides those listed in the table. [a]For all analyses predicting Smarter Balanced Writing Scale Score the teacher-level predictors were aggregated at the school level. *$p < .05$. **$p < .01$. ***$p < .001$.

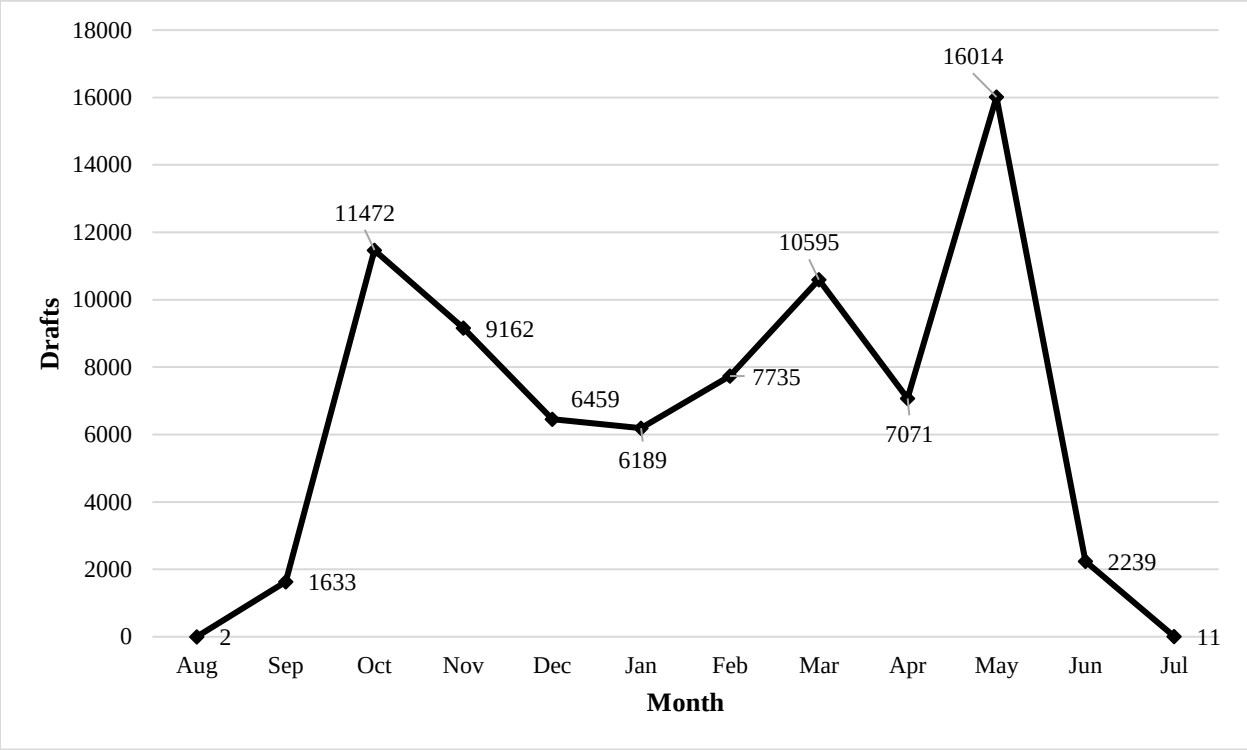


Figure 1. Total drafts submitted to MI Write per month by students in Grades 3–5 during SY 2017–18. Total *N* = 78,582 drafts.

Appendix – Screenshots of MI Write

Writing Analysis

Use this prompt to write an argumentative essay of your choice.

Show Highlights

Sleep

Students receive a lot of homework. When they home from school, it is hard to juggle the homework with other Sactivties. There needs to be time to eat dinner and often like to talk on the phone or the computer about Sthere day. This can be done on the weekend but that would also be cutting into their social lives. They do not have time to socialize with friends throughout the school day due to their homework. A lot of teenagers have trouble sleeping at night and do not go to bed until very late. They come to school at such an early hour Gand an barely function. They do not absorb anything when they are that tired. Even just a little extra time to sleep would be helpful to them. The point of high school is to prepare kids to go out into the world, so why not have Sthere hours match normal working hours? Starting Sridiclously early in the morning just forces Severone to get up early, including Gthe teachers. they have to get up even earlier than the students because they will have to get there first. A later start is a break for everyone.

Cited sources:
No sources entered.

Submissions

Latest Draft
2nd Draft
➜ 1st Draft

Add Comment
Mark Spelling
Mark Grammar

Peer Review

Not Requested

Word count: 198

Scores

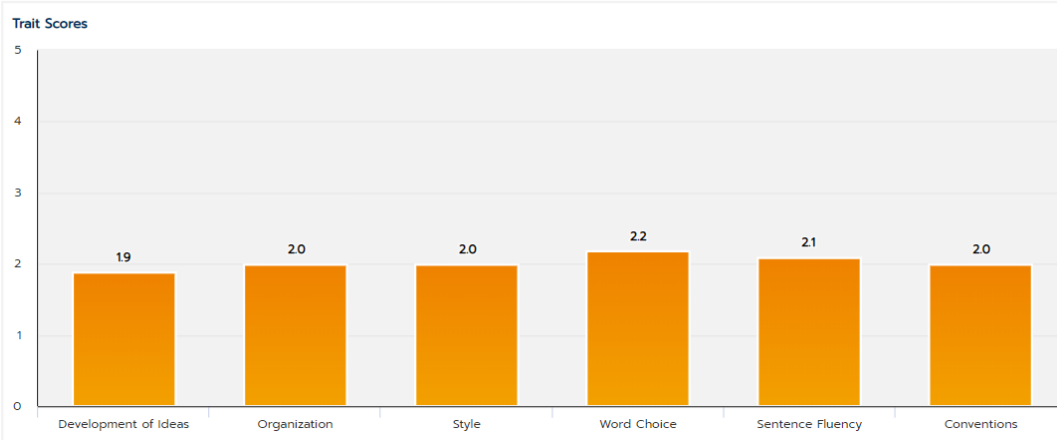
This essay's total score is 12.2

Percentile Rank is **21**
Stanine is **3**

Click here to learn more about percentile ranks and stanines.

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.



Style

★★☆☆☆

You scored 2.0 out of 5

Evaluation

- Does your essay give a sense of what you really think about the topic?
- Does your essay have conviction or individuality?
- Have you written your essay with your task, purpose, and audience in mind?
- Is your essay written in a formal style?

Lessons

- Understanding Audience

Word Choice

★★☆☆☆

You scored 2.2 out of 5

Evaluation

- Are the words in your essay interesting and uncommon?
- Do you use a variety of words and phrases or do you repeat the same ones?
- Have you chosen the correct words to express your ideas and to make the meaning of each sentence clear?
- Have you chosen words that are specific?

Feedback

- Check your writing for places you can choose words that are more specific to make your writing more informative. Try adding strong verbs, specific nouns, adjectives, and adverbs.

Lessons

- Appositives

Sentence Fluency

★★☆☆☆

You scored 2.1 out of 5

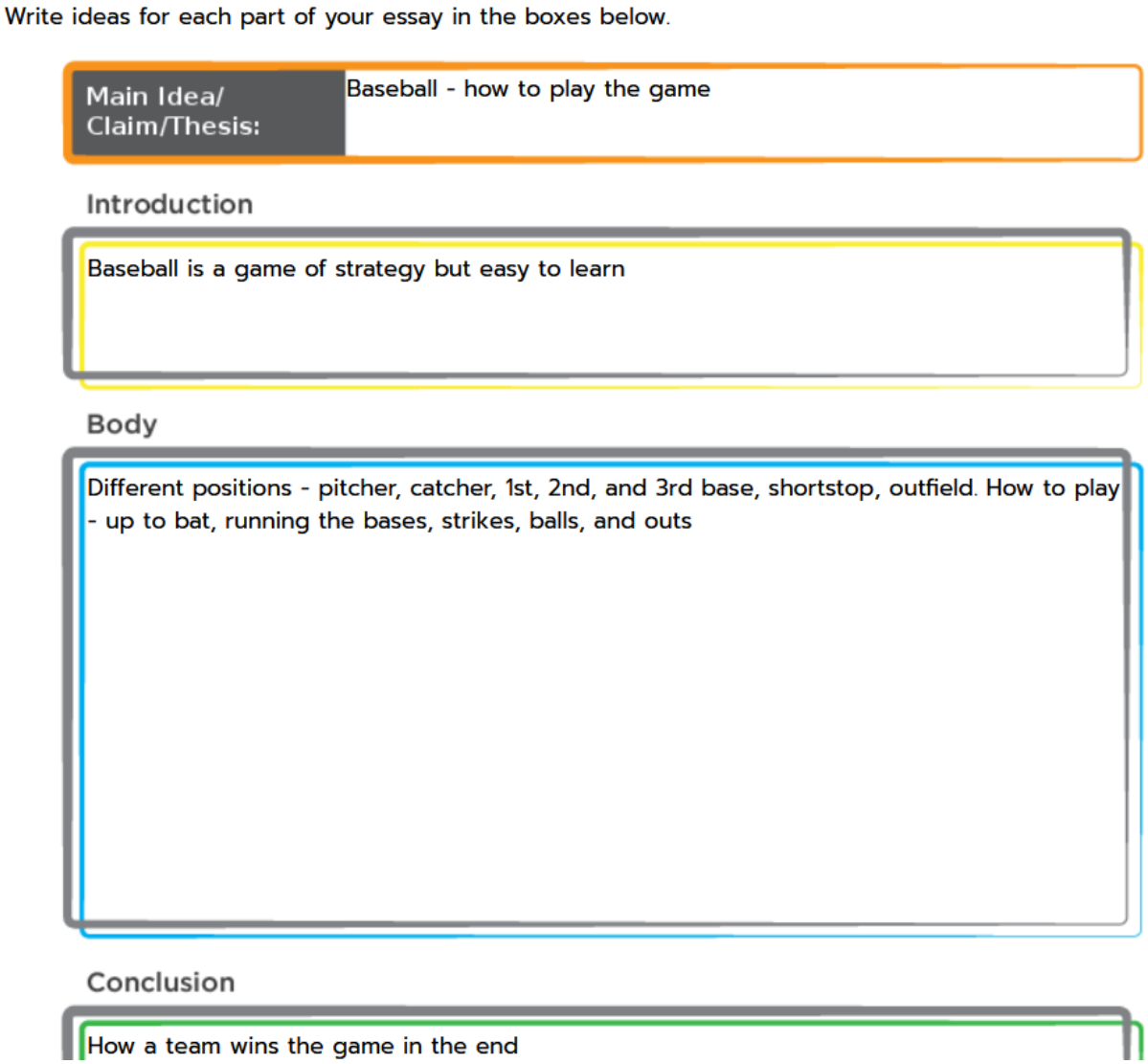
Evaluation

- Are your sentences easy to understand?
- Have you made sure that sentence fragments or awkward sentences do not disrupt the flow of your essay?
- Have you used different types of sentences instead of just simple sentences?

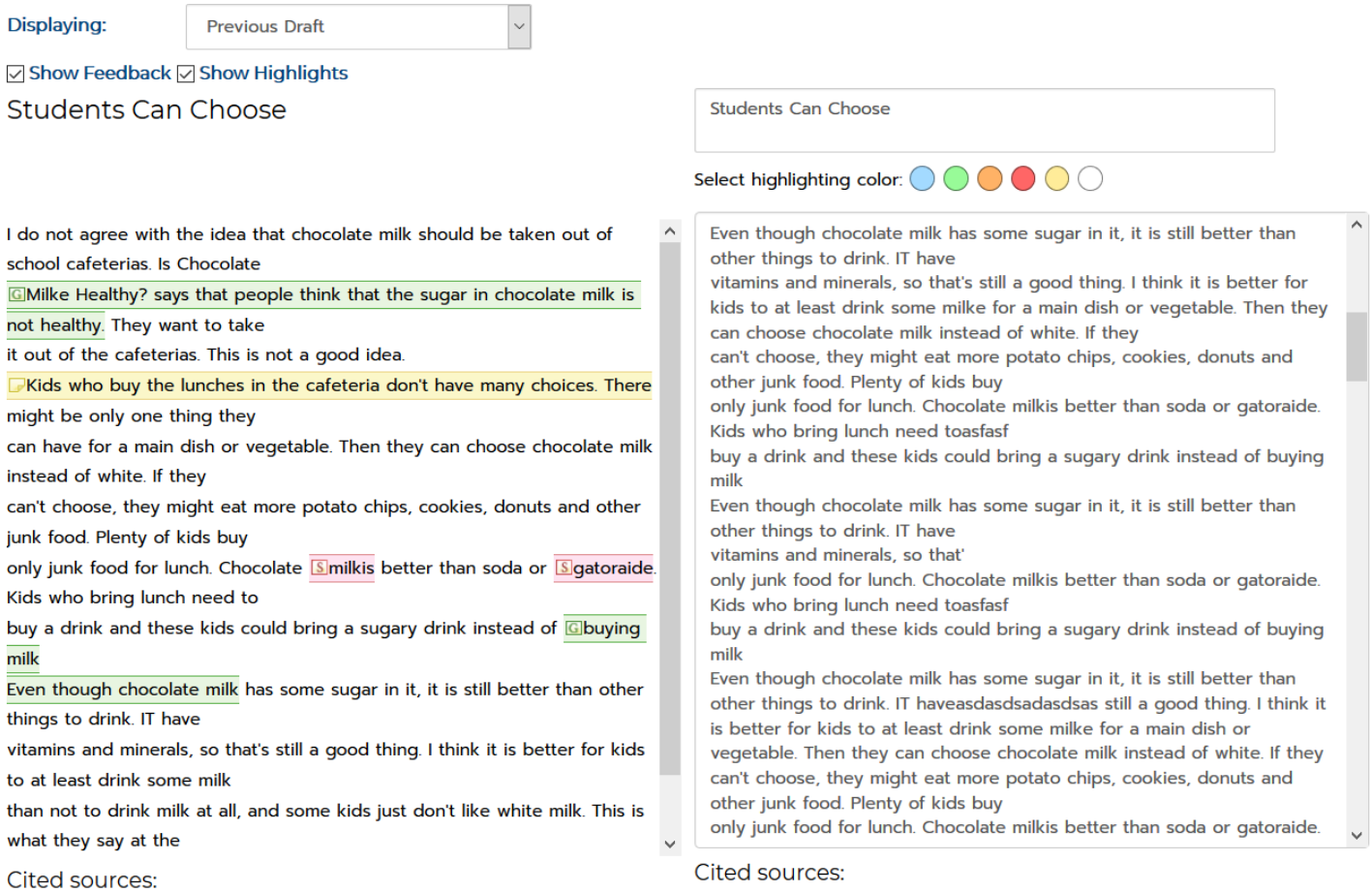
Lessons

- Making Sentences More Interesting

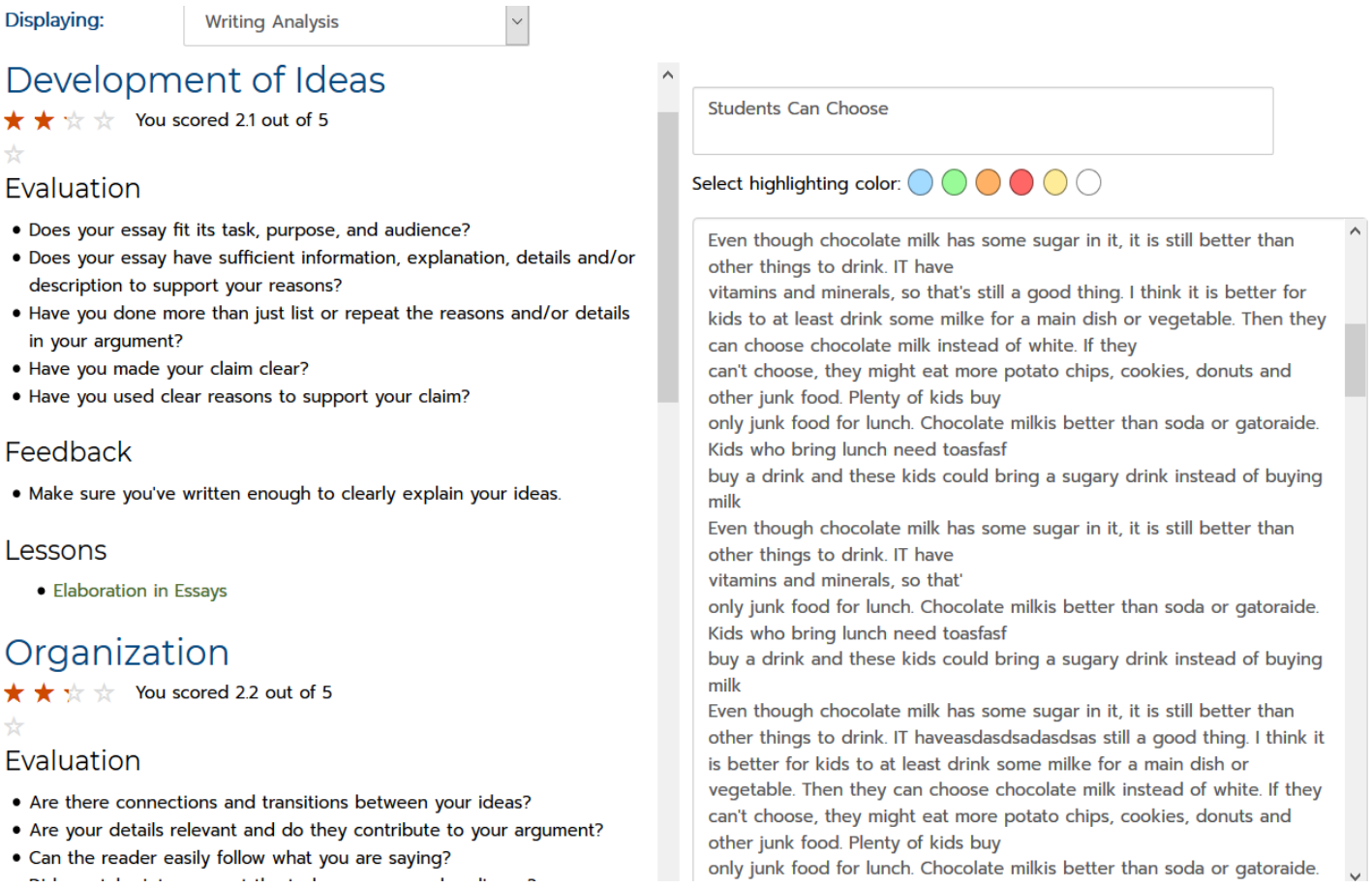
Example graphic organizer



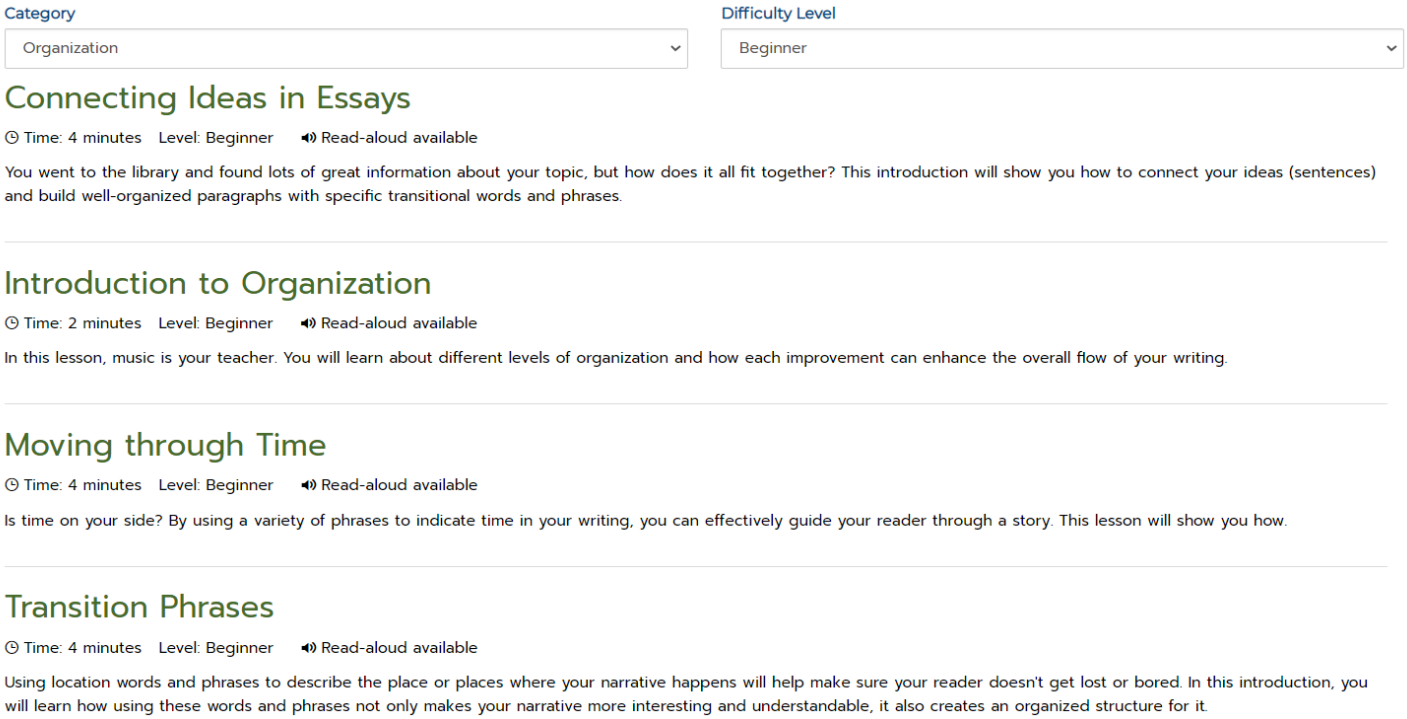
Split screen revision with spelling/grammar feedback



Split screen revision with writing analysis



Sample Lessons



Student portfolio

Student Averages: Trait and Total Scores

The report below displays the average score for each writing trait, total score, and teacher-scored textual evidence and content accuracy categories. The averages are calculated using the score from the most recent draft of each essay. To view the Score Report for an essay, click on the date.

Download

Date	Prompt	Drafts	Dev	Org	Style	WC	Sent	Conv	Total	% Rank	Stanine	Text	Cont
7/24/20 7:54 AM	A Special Memory	5	4.3	4.4	4.2	4.4	3.8	3.9	25.0	94	8	-	-
3/26/20 4:13 PM	Student Choice Informative/Explanatory	2	1.0	1.0	1.0	1.0	1.0	1.2	6.2	4	2	-	-
3/26/20 4:11 PM	Student Choice Narrative	1	2.7	2.8	2.8	3.1	2.9	3.0	17.3	40	5	-	-
3/26/20 3:36 PM	Student Choice Opinion	1	2.6	2.5	2.7	2.8	2.7	1.2	14.5	30	4	-	-
3/17/20 2:30 PM	A Favorite Activity	1	1.2	1.2	1.3	1.3	1.2	1.2	7.4	4	2	-	-
3/17/20 12:35 PM	All About Dirt	2	1.6	1.7	1.9	1.9	2.0	2.1	11.2	8	2	-	-
3/04/20 8:47 AM	All About Dirt	2	1.0	1.0	1.0	1.0	1.2	1.3	6.5	4	2	-	-
2/13/20 2:46 PM	Student Choice Opinion	4	2.8	2.6	3.1	3.2	2.9	3.2	17.8	62	6	-	-
8/22/19 8:56 AM	Chocolate Milk in Schools	3	2.1	2.2	2.5	2.5	2.5	2.6	14.4	29	4	1.0	2.0
4/25/19 1:52 PM	Chocolate Milk in Schools	5	2.7	2.7	3.1	3.1	3.1	3.3	18.0	62	6	-	-
1/13/19 3:12 PM	Student Choice Informative/Explanatory	3	4.1	4.3	4.3	4.3	4.4	4.5	25.9	95	8	-	-
12/13/18 9:35 PM	A Favorite Activity	2	3.9	4.1	4.2	4.3	4.4	4.5	25.4	94	8	-	-
8/21/18 8:37 PM	Chocolate Milk in Schools	1	2.5	2.5	3.0	2.9	2.9	3.1	16.9	37	4	2.0	1.0
8/21/18 3:02 PM	A Favorite Activity	3	4.0	4.1	4.2	4.3	4.4	4.5	25.5	95	8	-	-
1/30/17 12:07 PM	All About Dirt	1	1.0	1.0	1.0	1.0	1.0	1.0	6.0	4	2	-	-