International Society of the Learning Sciences

The Impact of Middle School Students' Writing Quality on the Accuracy of the Automated Assessment of Science Content

Dana Gnesdilow[1], Indrani Dey[1], Diane Gengler[1], Linda Malkin[1], Sadhana Puntambekar[1]
gnesdilow@wisc.edu, idey2@wisc.edu, dkgengler@wisc.edu, lmmalkin@wisc.edu,
puntambekar@education.wisc.edu
[1]University of Wisconsin – Madison
Rebecca J. Passonneau[2], ChanMin Kim[2]
rjp49@psu.edu, cmk604@psu.edu
[2]Pennsylvania State University

Abstract: Helping students learn how to write is essential. However, students have few opportunities to develop this skill, since giving timely feedback is difficult for teachers. AI applications can provide quick feedback on students' writing. But, ensuring accurate assessment can be challenging, since students' writing quality can vary. We examined the impact of students' writing quality on the error rate of our natural language processing (NLP) system when assessing scientific content in initial and revised design essays. We also explored whether aspects of writing quality were linked to the number of NLP errors. Despite finding that students' revised essays were significantly different from their initial essays in a few ways, our NLP systems' accuracy was similar. Further, our multiple regression analyses showed, overall, that students' writing quality did not impact our NLP systems' accuracy. This is promising in terms of ensuring students with different writing skills get similarly accurate feedback.

Introduction

One of the most important goals in education is to help students develop writing skills in all subject areas (Chen et al., 2022; Nunes et al., 2022). Despite this goal, most students in the United States struggle to write using standard English conventions to ensure clear and easy reading (Graham et al., 2015). For example, only about one quarter of 8th and 12th grade students performed at or above the proficient level for writing on the most recently reported National Assessment of Educational Progress assessments in 2011 ((NCES, 2012) Results from the NEAP assessment in 2017 could not be reported due to confounding issues in measurement). While research about the nature of student errors in adolescent writing is rare, several errors that impact writing quality have been identified, such as: misspellings, capitalization errors, incorrect or missing punctuation, inappropriate or missing commas, missing words, use of words out of context, using the wrong form of a word, and using the wrong verb tense (Wilcox et al., 2014). Further, students generally have more difficulties and make more errors when writing in subject-area versus in Language Arts classes (Lawrence et al., 2013; Wilcox et al., 2014).

The best way to learn to write is to engage in writing in different genres, where ideas are revised and refined over time (Graham, 2019, Kellogg & Whiteford, 2009). However, while students may have opportunities to write and revise narrative or persuasive pieces in language arts classes, they are rarely asked to write expository texts in other subjects (Graham et al., 2014) and writing instruction is generally insufficient (Graham, 2019). Further, students seldom receive formative feedback on their writing (Applebee & Langer, 2011) or make revisions. This is often due to the time it takes teachers to read and provide feedback (Chen et al., 2022; Zhai & Ma, 2022). Thus, although we expect students to have good practices in writing and making revisions, they may not have the opportunity to do so in an environment where they are scaffolded to develop these skills. With the rise of and improvements in artificial intelligence (AI), natural language processing (NLP) techniques, and large language models, technology is being developed to provide automated assessment and address this challenge to provide timely feedback on students' writing in a variety of genres (Gerard & Linn, 2022; Liu et al., 2016; Nunes et al., 2022) as well as for delivering rapid results in high stakes testing (Shermis, 2020; Shin & Gierl, 2021). Automated assessment and feedback systems can support students to improve the quality of their writing (Roscoe & McNamara, 2013; Zhai & Ma, 2023) or assess the content of students' writing to provide feedback about the important ideas they have or have not included (e.g., Kim & McCarthy, 2021; Madnani, et al., 2017).

While automated assessment and feedback on students' writing can provide students with quick formative feedback to identify areas for revision and improvement, the accuracy of these models is essential yet challenging (Liu et al., 2016; Nunes et al., 2022). Several aspects of students' writing might affect the accuracy of an automated assessment system, such as the length of the writing, sentence structure, cohesion between sentences and within an entire text, and other aspects of writing quality. For example, Crossley and McNamara (2016) found that higher levels of cohesion within a text indicated better quality writing and coherence. Relatedly,

Crossley et al. (2014) found that there may be multiple profiles for high-quality writing, and some high-quality writing may have lower syntactic and lexical complexity.

While studies have explored the accuracy of automated assessment (Bai & Hu, 2017) or scoring (Liu et al., 2016) systems, few have explored the extent to which the qualities of students' writing might impact the accuracy of the automated assessment of content. Since many middle school students struggle to write effectively, including students with individualized education programs and non-native speakers and writers of English, we wanted to understand if there were differences in the accuracy of automated assessment of students' design essays (described later) depending on the quality of students' writing. It is essential to ensure that students with a variety of backgrounds and skills can equally benefit from AI applications in educational settings. The purpose of this study was to investigate how middle school students' writing quality affected our automated assessment system's accuracy in identifying and providing feedback on the science ideas students included in their design essays during a roller coaster unit. The NLP assessment and feedback system we used is called PyrEval (Gao et al., 2018). We used it to assess the science content and relationships in middle school students' science explanations in design essays. Our research questions were: 1) How does the quality of students' writing affect PyrEval's accuracy (as assessed by a human) in identifying whether science ideas are present in their initial and revised essays? i.e., how might the length, inclusion of data, sentence complexity, cohesion, etc., of an essay impact the accuracy of an NLP system? and 2) Which aspects of writing quality, if any, are associated with fewer or more errors?

Methods

Participants

Three 8th-grade science teachers and their 238 consenting students from two semi-rural public-school districts in the midwestern United States participated in this study. Teacher 1 worked with 96 students in one school district and Teachers 2 and 3 worked in another district with 80 and 62 students, respectively. Both schools served 7th and 8th-grade students who were mostly white (about 15-20% non-white). All middle school students attend science class, including students identified as having disabilities and students whose first language is not English; thus, students vary greatly in their academic performance and writing skills in the same class. Prior to the implementation, we collaborated with the teachers to solicit their feedback on our curricula and technologies, as well as provided professional development for implementing design- and inquiry-based units and how to support students' use of PyrEval's automated feedback to help them revise their science writing.

Instructional context

The context of this research was a design-based roller coaster unit to help students learn about physics concepts and relationships about motion, forces, and energy. The unit took place over approximately fifteen, 50-minute instructional periods as a part of students' normal science classes. Students were challenged to design a rollercoaster that was both fun and safe for an amusement park whose attendance was waning. Students engaged in cycles of inquiry to learn about the science concepts and relationships they needed to solve the challenge. They first conducted background research on the relationships between height, mass, and energy. After this, they performed three experiments using a roller coaster simulation to explore a) how initial drop height impacts the amount of energy the car will have for the ride, b) the relationship between the initial drop height and subsequent hills to ensure the car can complete the ride, and c) the impact of the car's mass on the amount of energy available. Finally, they wrote an essay to explain their roller coaster design. Students submitted their essays to PyrEval, to get automated assessment and feedback on their work. Students then revised their essays using this feedback after engaging in a peer editing activity to brainstorm how to revise their writing using the feedback. Students received automated feedback on their revised essay for reflection as well. All students' work on the roller coaster unit, including use of the simulation and submission of essays to PyrEval, was accomplished in a digital notebook we developed. The online digital notebook provided structure and scaffolding for students to engage in the activities as well as providing a place for students to keep track of their ideas.

Data sources and analyses

We used several data sources. These are summarized in Table 1 and explained in detail in each section.

Students' design essays and total CU scores

While 236 students submitted an initial and revised design essay from the three teachers' classes, the data for our analyses are from 60 randomly selected students (n = 20 per teacher). We gave students the following instructions for writing their design essays after completing their investigations: "*Explain the science behind why your team's current roller coaster design will be exciting and make it to the end of the ride without stopping. Include data*

from your trials to justify your ideas. Make sure you write in clear and complete sentences". The teachers introduced the writing activity and provided students with prompts about: a) general tips for writing clear and concise essays and b) the science relationships they should include in their essays.

Our team built upon PyrEval to automatically assess and provide feedback on the content of science ideas and relationships that students included in their roller coaster design essays. PyrEval uses a wise-crowd model to identify weighted vectors of key content ideas and relationships from a small sample of reference responses (Gao et al., 2018), called content units (CUs). In our case, we used prior middle school students' roller coaster design essays for our wise-crowd model. Highly weighted CUs are more important to include in a text. For our roller coaster unit, the highly weighted CUs students should have included in their design essays were: 1) the greater the height of the initial drop the greater the potential energy at the top, 2) the inverse relationship between potential and kinetic energy as the car moves up and down the track, 3) total energy in a roller coaster system (without friction) equals the potential plus kinetic energies at any point on the track, 4) the Law of Conservation of Energy, 5) the initial drop height must be higher than subsequent hills for the car to have enough energy to finish the ride, and 6) a greater car mass means greater energy for the ride. These CUs were also highly aligned with the physics ideas and relationships students should have learned during the roller coaster unit. After writing and submitting their essays, PyrEval parsed each essay into propositions and identified whether each CU was present or not. It then produced a vector score for each CU; a 1 indicated that a particular CU was detected, whereas a score of 0 meant it was not. The feedback produced from the vector scores was presented to students in a checklist format, indicating which of the 6 CUs PyrEval identified in the essay, as well as the ones that were not detected. Students used this feedback to revise. After submitting their revised essay, they received feedback from PyrEval again to reflect further.

Vector scores for researchers were recorded as the following: [1,0,1,1,0,0]. In this example, PyrEval detected that CUs 1, 3, and 4 were present in a student's essay. Students could earn up to 6 CUs for a total CU score. We used the total CU scores from each student's initial and revised essays to assess whether their CU scores increased in their revision. These vector scores were also used to assess PyrEval's accuracy.

Table 1
Overview of data sources

Data	Description
Total CU scores	Total number of content units (CUs) identified by PyrEval per essay, 6 = max CU score
Human coding of PyrEval's accuracy	The number of PyrEval errors in assessing and providing vector scores for students' initial and revised essays as assessed by human coders
5 Measures of writing quality from TERA	Writing quality metrics assessed by Text Ease and Readability Assessor (TERA): narrativity, syntactic simplicity, concreteness, and referential and deep cohesion
Average sentence length	The average number of words per sentence for each essay
Length of essay	Number of words for each essay
Inclusion of data	Binary code indicating if students included data or not in each essay

Human coding of essays to assess PyrEval's accuracy

Two authors independently read and coded 20% of the 60 students' initial and revised essays (120 total essays) and compared their codes to PyrEval's vector scores to assess the number of errors PyrEval made in evaluating the presence or absence of the 6 key CUs in students' design essays (discussed previously). These errors could have been either false positive or negative errors. False positives were recorded when PyrEval determined that a CU was present in a students' writing when it was not. Oppositely, a false negative error was recorded when PyrEval did not detect that a particular CU was present, when, in fact, it was. We calculated a Cohen's Kappa statistic to ensure the coding between raters was reliable. We achieved substantial agreement (Kappa = .768) (Stemler, 2001). Discrepancies between raters were discussed and then one of the authors coded the remainder of the data. We then counted the number of errors PyrEval made on each essay for each student. We added these errors up for each set of essays and calculated the total percent of errors for the initial and then the revised essays.

Five measures of writing quality as measured by TERA

To understand how the quality of students' writing may have impacted PyrEval's accuracy, we used the Text Ease and Readability Assessor (TERA) to assess the writing quality of students' essays. TERA is a tool that uses Coh-Metrix (McNamara et al., 2014) to analyze complex features of text to provide measures of text "easability" and readability. While TERA is often used to select grade-appropriate texts for students, studies examining students'

writing using TERA have been conducted, including assessing the writing quality of non-native English speakers (e.g., Allagui, 2019; Msuya, 2017). Some research has identified that higher levels of cohesion within a text indicate better quality writing and coherence (Crossley & McNamara, 2016; Crossley et al., 2014). Crossley, et al. (2014) also found that some high-quality essays had lower syntactic and lexical complexity, demonstrating that there may be multiple profiles for high-quality essays. Although TERA analyzes multiple parameters of a text's readability, it presents its assessment as five components: 1) *narrativity*, how story-like a text is, greater narrativity means a text is easier to read; 2) *syntactic simplicity*, the number of words, clauses, and verb location in a sentence, a lower syntactic simplicity score indicates harder to read text; 3) *word concreteness*, the extent to which words are more or less abstract, use of more concrete words results in easier reading; 4) *referential cohesion*, overlap of words / ideas between sentences, higher scores indicate easier reading; and 5) *deep cohesion*, the extent to which ideas and events are connected in an entire text, higher scores means more connective words tie the text together, making it easier to read (see the T.E.R.A. website for a more detailed explanation of these text features and how they are measured). We believed that TERA's simple metrics of text quality could be used to help us understand whether the quality of students' writing could affect PyrEval's ability to recognize content units in essays. Each essay in our dataset was fully deidentified and assessed by TERA. We recorded the five component outputs for each essay, and then used them as independent variables in a multiple regression analysis.

Other qualities of students' writing that could impact PyrEval's accuracy

Average sentence and essay length: In prior analyses, we found that essays with an average of more than 25 words per sentence decreased PyrEval's ability to identify CUs (Puntambekar et al., 2023). Long sentences are difficult to read (Matthews & Folivi, 2023), and, as mentioned earlier, middle school students commonly fail to include punctuation in their writing. We wanted to see whether our improved PyrEval model mitigated this issue. We also noticed that very long essays contained repetitive ideas and sometimes crashed the system which might also impact PyrEval's accuracy. We wanted to rule out whether sentence or essay length impacted PyrEval's ability to identify CUs in students' writing. Thus, we calculated the average number of words per sentence and total number of words for each essay. These were then added as independent variables in our regression model.

Inclusion of data in essays: As part of the roller coaster design essay instructions, we prompted students to include data from their virtual experiments to support their science explanations and justify their roller coaster designs. Using data to support scientific explanations is a key science and engineering practice identified in the Framework for K-12 Science Education (NRC, 2012). While many students did include data, many did not. However, in examining essays that did include data, we noticed that students sometimes inserted decimal points in the numbers or included periods when abbreviating units of measurement, e.g., writing "*5.00 m.".* We wondered if the extra periods influenced how PyrEval parsed the essay into propositions, potentially impacting its accuracy. Furthermore, PyrEval does not evaluate numbers, which could have affected how PyrEval interpreted sentences with data. To explore whether students' inclusion of data may have influenced PyrEval's accuracy, we coded each essay as either "1", included data, or "0", did not include data. This was the final independent variable in our multiple regression analysis.

Results

As students' writing quality may differ between the original and revised essays, we first compared the quality of students' writing between these two datasets to provide context for the interpretation of our regression analyses.

Nature of students' writing in the initial and revised essays

Changes in total CU scores between initial and revised essays

We examined whether PyrEval identified more CUs in students' revised versus initial essays. We ran a repeated measures analysis using students' total CU scores as the dependent variable. We found that PyrEval detected significantly more CUs in students' revised essays than in their initial essays ($F_{(1,59)} = 48.69$, $p < .001$, $\eta p^2 = .452$). See Table 2 for descriptives and repeated-measures significance.

Changes in the quality of students' writing between initial and revised essays

In addition to examining whether students' total CU scores were different between their initial and revised essays, we also wanted to know if there were differences in their writing quality to provide context for interpreting our regression analyses (described below). We performed a repeated measures analysis for each aspect of writing quality between the initial and revised essays: narrativity, syntactic simplicity, concreteness, referential cohesion, deep cohesion, total number of words, average words per sentence, and inclusion of data. We found that students'

writing was significantly less concrete in their revised essays than their initial essays ($F_{(1,59)} = 28.71$, $p < .001$, $\eta p^2 = .327$), indicating the inclusion of more abstract terms, like energy. Students also included significantly more words, and more students included data in their revised versus initial essays (($F_{(1,59)} = 66.24$, $p < .001$, $\eta p^2 = .529$) and ($F_{(1,59)} = 12.21$, $p < .001$, $\eta p^2 = .171$), respectively) (Table 2). We found no other significant differences between the initial and revised essays in terms of writing quality.

Table 2
Means (SDs) for the eight writing quality categories and repeated measures significance

	Narrativity	Syntactic simplicity	Concrete-ness	Referential cohesion	Deep cohesion	# of words	Average words / sentence	Included data	Total CU score
Initial Essay	61.87 (16.61)	29.32 (18.48)	32.77 (20.91)	93.35 (12.81)	74.97 (27.28)	358.58 (154.53)	22.25 (7.11)	0.68 (0.47)	4.25 (1.41)
Revised Essay	61.45 (16.10)	30.53 (19.41)	*25.28 (17.36)	95.37 (9.71)	77.83 (25.05)	*436.87 (168.56)	20.79 (5.41)	*0.88 (0.32)	*5.22 (1.01)

Repeated measures significance: *$p < .001$

Results of human coding of PyrEval's accuracy

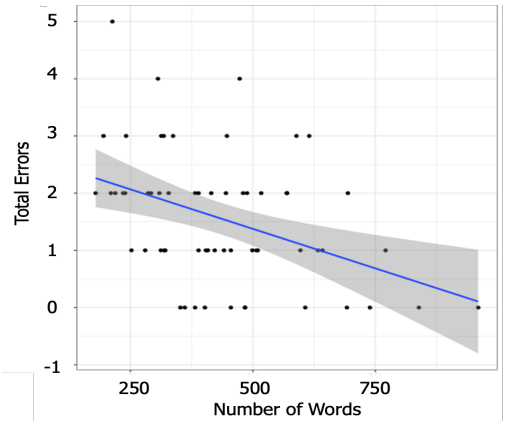
Based on the idea that there were differences in students' writing quality between the initial and revised essays, we also explored whether there were differences in the percentage of errors PyrEval made in assessing students' initial and revised essays. We added the total number of errors in each set of essays and divided it by 360 (60 initial (or revised essays) x 6 CUs per essay) and multiplied by 100. We found that the percentage of errors PyrEval made was similar for the initial and revised essays, about a 25% and 26% error rate, respectively.

The impact of students' writing quality on PyrEval's accuracy

To determine whether the eight qualities of students' writing we quantified were predictive of PyrEval's error rates, we conducted two multiple linear regression analyses, the first on the initial essay and the second on the revised essay. Each regression analysis used the total number of errors PyrEval made as the dependent outcome and the eight different aspects of students' writing (narrativity, syntactic simplicity, concreteness, referential and deep cohesion, number of words, average words per sentence, and inclusion of data) as predictors.

We found no significant association between the number of errors PyrEval made and the quality of students' writing on the initial essay ($F_{(8,51)} = 0.672$, $p = 0.714$), with an R^2 of 0.095. Similarly, the regression equation was not significant for the association between the eight writing quality metrics and PyrEval's error rate on the revised essays ($F_{(8,51)} = 1.635$, $p = 0.138$), with an R^2 of 0.204 . Even though the overall regression equation for the revised essay was not significant, we found that a one-unit increase in the number of words in an essay was associated with a .0025 unit decrease in the total number of errors made by PyrEval ($p = 0.009$) (see Figure 2). This means that PyrEval made slightly fewer errors when the revised essays were longer.

Figure 2
Association of number of PyrEval errors and length of revised essay



Discussion

The goal of this study was to understand whether the quality of middle school students' writing impacted our NLP system's ability to accurately assess the science ideas students included in their design essays and if particular aspects of students' writing quality were associated with fewer or more errors. Thus, we explored how eight aspects of writing quality may have affected PyrEval's ability to identify whether science ideas were present or not in students' initial and revised design essays through performing multiple linear regression analyses. To interpret any potential differences in our regression analysis we also explored whether there were differences in each of the eight aspects of writing quality between students' initial and revised essays, as well as PyrEval's accuracy rate on these two sets of essays. We found that even though students' revised essays included significantly more abstract words, words overall, content units, and more students included data than in their initial essays, PyrEval's error rate on both essays were similar.

Our regression analyses helped us identify that essay writing quality was not significantly associated with PyrEval's accuracy for our 8th grade student population. However, the model for the revised essays indicated that a greater number of words in the essays may be associated with a very slight increase in PyrEval's accuracy. Our motivation for the inclusion of the number of words in our analysis was based on the idea that longer essays sometimes crashed the PyrEval system and there could be a lot of repetition of ideas that might negatively impact PyrEvals's accuracy. However, perhaps this increase in accuracy, while very small, may have occurred because students reiterated their ideas or revised with more precise wording, giving PyrEval more opportunities to "catch" the presence of content units. Anecdotally, in our analyses of the accuracy of PyrEval, our team noticed that students who wrote longer essays tended to explain the content units in multiple, but slightly different, ways or added more specific language, such as adding the word "total" to energy, when previously they only wrote energy.

We find these results to be encouraging. Even though we surmised that extra periods when data was included in essay might affect PyrEval's performance, we found no evidence for this. The inclusion of data as evidence in science writing is an essential science and engineering practice (NRC, 2012). It is important to know that students' engagement in this practice did not influence the accuracy of the automated assessment and the feedback they received from PyrEval as compared to students who do not include data. Our findings indicate that PyrEval detected science ideas in students' essays with similar accuracy, regardless of the quality of their writing. This may mitigate potential biases for students who struggle with expressing their ideas in writing. Finally, students benefited from receiving automated feedback on the science content and having the opportunity to reflect upon and revise their writing, since they included a significantly higher number of science content units in their revised essays.

Implications, limitations, and future research

Many middle school students struggle to write using standard English conventions and their writing commonly contains a variety of errors. This is especially true when they write in subjects other than Language Arts (Wilcox et al., 2014). For example, scientific writing is complex, where students need to construct explanations using evidence from their experiments to support their claims (Berland & Reiser, 2009). To further compound this problem, students are rarely asked to write and revise expository texts based on feedback (Graham et al., 2014), which would provide them with much needed reflection and practice to improve in these skills and learn content and relationships associated with the topic on which they are writing. With the development of AI models, students' writing can be automatically assessed, and feedback can be provided quickly, mitigating challenges teachers may have in providing timely and actionable feedback. However, its utility may be diminished if students do not find it to be accurate (Bai & Hu, 2017); therefore, it is essential that researchers work to ensure that the assessment and feedback on students' writing is accurate and, thus, should thoroughly explore the factors that may impact an AI system's accuracy. This is not only important because we want students to reflect on accurate constructive feedback to make appropriate revisions and learn, but we also want to make sure that the accuracy of the automated assessment and feedback is similar for students with a variety of skills and backgrounds.

Artificial intelligence applications in education have the potential to ensure that students get the timely, formative feedback they need to develop new skills and learn in a variety of activities and subject areas. Even though PyrEval's assessment of the content of students' writing and feedback was wrong about 25% of the time, students still benefited from using the feedback to improve their writing. We are working on ways to increase the accuracy of PyrEval and predict that receiving more accurate feedback will result in even better outcomes for students. While we have investigated several aspects of writing quality in this study, there are a multitude of ways that students' writing may vary that we did not examine, which could impact PyrEval's accuracy. Further, the results of this study are based on a relatively small sample of students from only three teachers' classes from semi-rural school districts. Even though we found that PyrEval was equally accurate despite essay writing quality, it

was trained on a similar population of student essays. This means PyrEval may perform more or less effectively when providing feedback on writing exhibiting novel or different conventions and patterns. Machine learning models can be biased depending on the algorithm with which they were trained (Fazelpour & Danks, 2021; Sun et al., 2020). Within the context of students' writing, non-native English speakers and students from minority groups may be less proficient and may not articulate ideas as clearly in their writing (Allagui, 2019; Crossley & McNamara, 2011; NCES, 2012). While researchers are developing methods to detect and address bias, including monitoring and evaluating these models to be fair and non-discriminatory for different groups (Holstein et al., 2018), or developing more nuanced fairness metrics (Binns, 2018), much more work needs to be done in this area.

In future work, our plan is to improve PyrEval's accuracy as well as explore other factors that may negatively impact its identification of important content units in students' writing. Further, we have been integrating classroom participatory structures and routines to complement AI generated assessments (Puntambekar et al., 2024) to address gaps in equity, and provide more support for students who might struggle with writing and in understanding automated assessment. We are also exploring how AI generated assessment can be utilized within a distributed scaffolding system (Puntambekar, 2022), where teachers and students work together to critically evaluate and effectively integrate AI assessment in meaningful ways. Additionally, we have been working with teachers to help them understand both the benefits and inaccuracies of AI generated assessments, so that they can help their students interpret automated feedback and revise their writing accordingly.

References

Allagui, B. (2019). Investigating the quality of argument structure in first-year university writing. *English language teaching research in the Middle East and North Africa: Multiple perspectives*, 173-196.

Applebee, A. N., & Langer, J. A. (2011). " EJ" Extra: A Snapshot of Writing Instruction in Middle Schools and High Schools. *The English Journal*, *100*(6), 14-27.

Bai, L. & Hu, G. (2017). In the face of fallible AWE feedback: how do students respond? *Educational Psychology, 37*(1), 67-81.

Berland, L. K., & Reiser, B. J. (2009). Making sense of argumentation and explanation. *Science Education, 93*(1), 26-55.

Binns, R. (2018). Fairness in machine learning: lessons from political philosophy. In *Proceedings of the Conference on Fairness, Accountability, and Transparency on Machine Learning Research, 81* (pp. 149–159).

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

Crossley, S. A., & McNamara, D. S. (2011). Shared features of L2 writing: Intergroup homogeneity and text classification. *Journal of Second Language Writing, 20*(4), 271-285.

Crossley, S. A., Allen, L. K., & McNamara, D. S. (2014). A Multi-Dimensional analysis of essay writing: What linguistic features tell us about situational parameters and the effects of language functions on judgements of quality. In T. Berber Sardinha & M. Veirana Pinto (Eds.), *Multi-dimensional analysis, 25 years on. A tribute to Douglas Biber*, *(pp.* 197-237). John Benjamins Publishing Company.

Crossley, S. A., & McNamara, D. S. (2016). Say more and be more coherent: How text elaboration and cohesion can increase writing quality. *Journal of Writing Research*, *7*(3), 351-370.

Fazelpour, S., & Danks, D. (2021). Algorithmic bias: Senses, sources, solutions. *Philosophy Compass, 16*(8), e12760.

Gao, Y., Warner, A., & Passonneau, R. J. (2018, May). Pyreval: An automated method for summary content analysis. *In Proceedings of the eleventh International Conference on Language Resources and Evaluation* (LREC 2018). Miyazaki, Japan. European Language Resources Association (ELRA).

Gerard, L. & Linn, M., C. (2022). Computer-based guidance to support students' revision of their science explanations. *Computers & Education, 176*, 104351.

Graham, S. (2019). Changing how writing is taught. *Review of Research in Education*, *43*(1), 277-303.

Graham, S., Capizzi, A., Harris, K. R., Hebert, M., & Morphy, P. (2014). Teaching writing to middle school students: A national survey. *Reading and Writing*, *27*, 1015-1042.

Graham, S., Hebert, M., & Harris, K. R. (2015). Formative assessment and writing: A meta-analysis. *The Elementary School Journal*, *115*(4), 523-547.

Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., & Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need?. *In Proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1-16).

Kellogg, R. T., & Whiteford, A. P. (2009). Training advanced writing skills: The case for deliberate practice. *Educational Psychologist*, *44*(4), 250-266.

Kim, M. K., & McCarthy, K. S. (2021). Improving summary writing through formative feedback in a technology-enhanced learning environment. *Journal of Computer Assisted Learning*, *37*(3), 684-704.

Lawrence, J. F., Galloway, E. P., Yim, S., & Lin, A. (2013). Learning to write in middle school? Insights into adolescent writers' instructional experiences across content areas. *Journal of Adolescent & Adult Literacy*, *57*(2), 151-161.

Liu, O. L., Rios, J. A., Heilman, M., Gerard, L., & Linn, M. C. (2016). Validation of automated scoring of science assessments. *Journal of Research in Science Teaching, 53*(2), 215-233.

Madnani, N., Loukina, A., & Cahill, A. (2017, September). A large scale quantitative exploration of modeling strategies for content scoring. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 457-467).

Matthews, N., Folivi, F. (2023). Omit needless words: Sentence length perception. *PLoS One*, 18(2), e0282146.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

Msuya, E. A. (2017). Assessing text easibility of university students' EFL writing in Tanzania. *Journal for the Study of English Linguistics*, 5(1).

National Center for Educational Statistics (2012). The nation's report card: Writing 2011 national assessment of educational progress at grades 8 and 12. U.S. Department of Education. https://nces.ed.gov/nationsreportcard/pdf/main2011/2012470.pdf

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas.* National Academies Press.

Nunes, A., Cordeiro, C., Limpo, T., & Castro, S. L. (2022). Effectiveness of automated writing evaluation systems in school settings: A systematic review of studies from 2000 to 2020. *Journal of Computer Assisted Learning*, *38*(2), 599-620.

Puntambekar, S. (2022). Distributed scaffolding: Scaffolding students in classroom environments. *Educational Psychology Review*, *34*(1), 451-472.

Puntambekar, S., Dey, I., Gnesdilow, D., Passonneau, R. J., & Kim, C. (2023). Examining the effect of automated assessments and feedback on students' written science explanations. In Blikstein, P., Van Aalst, J., Kizito, R., & Brennan, K. (Eds.), *Building Knowledge and Sustaining our Community: Proceedings of the 17th International Conference of the Learning Sciences - ICLS 2023,* (pp. 1866-1867). Montreal, Canada: International Society of the Learning Sciences.

Puntambekar, S., Gnesdilow, D., Passonneau, R. J., & Kim, C., (2024). AI-human partnership to help students write science explanations. In press.

Roscoe, R. D., & McNamara, D. S. (2013). Writing Pal: Feasibility of an intelligent writing strategy tutor in the high school classroom. J*ournal of Educational Psychology, 105*(4), 1010.

Shermis, M. D. (2010). Automated essay scoring in a high stakes testing environment. In V.J. Shute and B.J. Becker (Eds.), *Innovative assessment for the 21st century: Supporting educational needs*, pp. 167-185.

Shin, J., & Gierl, M. J. (2021). More efficient processes for creating automated essay scoring frameworks: A demonstration of two algorithms. *Language Testing*, *38*(2), 247-272.

Stemler, S. (2001). An overview of content analysis. *Practical Assessment, Research & Evaluation*, *7*(17), 137-146.

Sun W., Nasraoui O, & Shafto P. (2020) Evolution and impact of bias in human and machine learning algorithm interaction. *PLoS ONE 15*(8): e0235502.

Wilcox, K. C., Yagelski, R., & Yu, F. (2014). The nature of error in adolescent student writing. *Reading and Writing*, *27*, 1073-1094.

Zhai, N., & Ma, X. (2022). Automated writing evaluation (AWE) feedback: A systematic investigation of college students' acceptance. *Computer Assisted Language Learning*, *35*(9), 2817-2842.

Acknowledgments

We are grateful to the students and teachers who participated in this research study. This research has been supported by a DRL grant from the National Science Foundation (Grant # 2010483).