International Society of the Learning Sciences

Factors Influencing Students' Perceptions of Automated Feedback and Their Impact on Revision

Indrani Dey, University of Wisconsin–Madison, idey2@wisc.edu
Dana Gnesdilow, University of Wisconsin–Madison, gnesdilow@wisc.edu
Riley Smith, University of Wisconsin–Madison, rasmith36@wisc.edu
ChanMin Kim, Pennsylvania State University, cmk604@psu.edu
Rebecca J. Passonneau, Pennsylvania State University, rjp49@psu.edu
Sadhana Puntambekar, University of Wisconsin–Madison, puntambekar@education.wisc.edu

Abstract: Automated feedback can provide students with timely information about their writing, but students' willingness to engage meaningfully with the feedback to revise their writing may be influenced by their perceptions of its usefulness. We explored the factors that may have influenced 339, 8th-grade students' perceptions of receiving automated feedback on their writing and whether their perceptions impacted their revisions and writing improvement. Using HLM and logistic regression analyses, we found that: 1) students with more positive perceptions of the automated feedback made revisions that resulted in significant improvements in their writing, and 2) students who received feedback indicating they included more important ideas in their essays had significantly higher perceptions of the usefulness of the feedback, but were significantly less likely to engage in substantive revisions. Implications and the importance of helping students evaluate and reflect on the feedback to make substantive revisions, no matter their initial feedback, are discussed.

Introduction

Learning to write is a constructive, complex process that requires students to use a variety of cognitive skills and knowledge to communicate effectively with others. This skill is developed through iterative experiences within cultural contexts and in interaction with others and tools (Beiki, 2022; Wilson et al., 2021a). However, limited time and resources are generally provided to support students' revision process so they can practice and develop their writing skills (Fleckenstein et al., 2023; Graham, 2019). In particular, it is difficult for teachers to provide the individual support students need (Collins et al., 2019) as well as provide the timely, comprehensive feedback needed for improvement (Chen et al., 2022; Zhai & Ma, 2023). With improvements in artificial intelligence and Natural Language Processing (NLP) techniques, systems for providing students with automated feedback on their writing have been developed that can provide students with constructive feedback in a timely manner (Liu et al., 2024; Roscoe et al., 2018; Wilson, 2021b).

Reviews or meta-analyses of studies investigating the impact of automatic writing evaluation and feedback systems in supporting students' writing have identified, by and large, that automated feedback systems can help students improve their writing (Fleckenstein et al., 2023; Fu et al., 2024; Nunes et al., 2022). However, it is important to understand the instructional contexts in which automated feedback systems are adopted, including students' readiness to use these systems (Delgado et al., 2024). Several studies have identified that students may not use automated feedback to engage in meaningful revisions of their writing (Dey et al., 2024; Puntambekar et al., 2023; Zhang, 2020). In these cases, some students may have difficulties using the feedback to revise, while others may simply make surface revisions or "game the system" to achieve a higher score, rather than engaging more meaningful improvements to their writing (Lottridge et al., 2021; Moore & MacArthur, 2016).

In their review, Fu et al. (2024) stressed that students' interactions with automated feedback are likely shaped by their emotions, cognition, and behavioral engagement, which, in turn, impacts the extent to which they engage with the revision process and improve. They further reported that students' learning and/or use of automated feedback may be influenced by their perceptions of the accuracy of the system, their prior knowledge, language or academic skills, or their teacher. In line with this, some research has identified that students who perceived automated feedback as more helpful or valuable were more accepting of using the technology and feedback to engage in more meaningful revisions (Nunes et al., 2022; Zhai & Ma, 2022). Conversely, other studies found that students revised using automated feedback regardless of their perceptions of it (Roscoe et al., 2017, 2018), or that there was no difference in perceptions between students who revised and those who did not (Zhu et al., 2020). Related research has identified that students' immediate automated feedback scores may impact their likelihood of engaging in revision, with some studies finding that students who received higher scores were more likely to revise and improve their writing (Zhu et al., 2017; 2020), and others finding the opposite (Moore & MacArthur, 2016). Wilson et al. (2024) investigated multiple factors that may predict students' perceptions of

automated writing feedback tools (AWF) and identified that less proficient writers found automated feedback to be more useful. They also found that while some individual factors predicted students' perceptions of automated feedback on their writing, others, such as gender, did not. While other studies investigating AWF tools collected gender information, they did not use it to examine differences in gender perceptions. Ofosu-Ampong's (2023) study, though not specific to AWF tools, suggested that gender impacts students' perceptions and use of AI-based tools in general, finding that males viewed AI-based tools more positively and used them more.

These conflicting findings indicate that students' perceptions and use of automated feedback are influenced by different factors, which are likely context dependent. Therefore, when designing and adopting new technologies to support students' writing, it is imperative to better understand how students' perceptions of AWF systems may impact how they utilize them (Fu et al., 2024; Wilson et al., 2024). This knowledge is vital for designing more effective systems and instructional methods to better support the development of students' writing skills. Thus far, the majority of studies examining students' perceptions have focused on the tertiary level and English writing and language classes (Nunes et al., 2022; Zhai & Ma, 2021). Moreover, few studies have examined students' perceptions of AWF tools to support their scientific explanation writing and revision, and even then, it may not be the primary focus (Zhu et al., 2020). Thus, the goal of this study was to explore the factors that may have influenced middle school students' perceptions of using automated feedback to improve the content of their science writing, as well as whether their perceptions impacted their engagement in revisions that might result in improvements to their scientific writing. Our research questions were:

1. What factors influence students' perceptions of automated feedback?
2. How do students' perceptions impact whether they revise and improve?

Methods

Participants

This study took place as a part of a multi-year design-based research project to develop an AI system that would provide students with feedback on the content of their scientific writing. Four-hundred and seventy-three 8th-grade students from six science teachers' classes at two middle schools in two different rural-fringe school districts in the midwestern U.S. consented to having their data collected for this study (School 1 = three teachers and 242 students, School 2 = three teachers and 231 students). From these participants, we were able to collect full data (students who submitted an initial essay, submitted a revised essay, completed the perceptions questionnaire, and took the content knowledge pre-test) from 339 students, which we used in our analyses.

Three of the teachers, one from School 1 and two from School 2, participated in all formal professional development meetings and implementations during all three years of the project and two teachers from School 1 used the automated feedback system in their classes every year of the project. One teacher in School 2 was new but participated in one formal professional development meeting. All teachers participated in informal professional development and email communications occurred regularly between the research team and the teachers in an attempt to increase the fidelity of implementation. We also provided a researcher-written teacher's guide to further support the teachers' implementations of the unit and use of the automated feedback in their classrooms.

Instructional context

The automated feedback system to support students' scientific writing, called PyrEval (Gao et al., 2018) (described later), was utilized within the context of an inquiry and design-based roller coaster unit aimed at helping students to learn about motion, forces, and energy while engaging meaningfully in science and engineering practices. The unit took place over approximately fifteen, 50-minute regular science classes in School 1, and about 20 classes in School 2. The roller coaster unit challenged students to design a roller coaster based on physics that was fun and safe to increase attendance at an amusement park. We designed the unit to engage students in multiple activities to help them learn about the physics behind roller coasters, including how the height of the initial drop and the mass of the car would impact the amount of energy there would be for the ride as well as about energy conservation and transformation. We then expected students to use what they learned and the data they collected during the unit to explain their roller coaster design in an essay that would be assessed by PyrEval, for which they would receive automated feedback. The unit unfolded as follows, students: 1) took a physics pretest; 2) were introduced to the roller coaster challenge; 3) conducted three simulated experiments and collected data and looked for patterns between variables; 4) wrote and submitted their initial roller coaster design essay to get feedback from PyrEval; 5) engaged in a peer review activity facilitated by the teacher to understand and reflect on how to use the automated feedback to revise; 6) received feedback from PyrEval and were given time to revise their essay based on the feedback; and 7) took an online perceptions questionnaire about the feedback they received. Students

used a digital notebook throughout the unit to structure and track their ideas, run simulated experiments, and submit their essays to PyrEval. Students then received automated feedback on the science ideas PyrEval identified in their essays within the digital notebook, regardless of writing quality (Gnesdilow et al., 2024).

PyrEval feedback and content units (CUs)

Our team refined the NLP models used by PyrEval to automatically assess the science content in students' roller coaster design essays and provide feedback. PyrEval uses a wise crowd model to identify weighted vectors of key content ideas and relationships (called content units (CUs)) within a small sample of reference responses (Gao et al., 2018). The more highly weighted a CU is, the more important it is to include in a text. For utilization in classroom settings with different target ideas, PyrEval can also be tuned to a specific main idea rubric (Singh et al., 2022). We identified six highly weighted CUs, or main ideas, students should have included in their design essays: 1) a higher initial drop means greater potential energy at the top, 2) as the car goes down the track the potential energy decreases and kinetic energy increases (and vice versa), 3) when there is no friction, the potential and kinetic energies at any point on the track add up to the total energy, 4) the Law of Conservation of Energy, 5) the initial drop must be higher than subsequent hills for the car to have enough energy for the ride, and 6) a roller coaster car with more mass means greater energy for the ride. These six CUs were also highly related to the key science ideas and relationships students would explore during the unit.

After an essay is submitted to PyrEval, the system parses it into propositions and identifies whether each CU is present or not. PyrEval produces a vector score indicating which CUs are present (denoted as a 1) or absent (denoted as a 0). An example vector score produced for researchers is [1,0,1,0,0,0]. This vector score indicates that PyrEval detected only CUs 1 and 3 in an essay. We presented the feedback from the vector scores as a table to students (Figure 1), listing each of the 6 CUs and indicating which ideas PyrEval identified in the essay (or not). Teachers facilitated a peer review activity, to guide students how to use the automated feedback to determine areas for revision, even if the feedback indicated they included particular ideas (CUs). Students also received feedback on their revised essays so they could further reflect on whether their revision improved their essays.

Figure 1

Sample feedback table generated by PyrEval for students, showing each content unit relationship and whether PyrEval detected it (green check mark) or not (orange question mark).

Feedback	
Height and Potential Energy	✓
Relation between Potential Energy and Kinetic Energy	✓
Total energy	✓
Energy transformation and Law of Conservation of Energy	?
Relation between initial drop and hill height	✓
Mass and energy	?

Data sources

AI feedback perceptions questionnaire

We created an AI feedback perceptions questionnaire consisting of eight Likert-style statements to understand students' perceptions about the usefulness of the automated feedback in helping them revise the science ideas in their design essays. The questionnaire had high Cronbach's alpha ($\alpha = 0.79$), suggesting strong internal consistency and reliability that the items likely measure the same underlying construct.

In the questionnaire, we asked students to state their level of agreement to the eight statements about the accuracy of automated feedback and how helpful or confusing they found the automated feedback for supporting their revisions (question stem and statements listed in Table 1). For each statement, students chose whether they strongly disagreed, disagreed, agreed, or strongly agreed. We then created a total perceptions survey score for each student. To do this we assigned different point values for each level of response. For positively stated items (1, 2, 3, and 5), strongly agree was assigned 4 points, agree was assigned 3, disagree was assigned 2, and strongly

disagree was assigned a 1. For negatively stated items (4, 6, 7 and 8), we assigned the opposite, e.g., strongly disagree as 4, etc. To create the total survey score, we summed students' responses to the six statements, which could range from 8 to 32 points. Higher perception scores indicated more positive perceptions about the automated feedback.

Table 1

Positive and negative Likert-style statements related to the question stem, "What are your thoughts about the automated feedback?", from the Perceptions Questionnaire

Questionnaire statements	Positive or negative item in analyses
1) The automated feedback was accurate.	Positive
2) The automated feedback helped me notice the main ideas I missed in my essay.	Positive
3) It was easy for me to understand the feedback and know what I needed to revise.	Positive
4) I understood the feedback, but I didn't know how to use it to revise my essay.	Negative
5) I agreed with the automated feedback.	Positive
6) The feedback was frustrating and did not help me revise.	Negative
7) I did not agree with the automated feedback.	Negative
8) The automated feedback was inaccurate.	Negative

Roller coaster design essays and CU scores

As mentioned previously, students wrote roller coaster design essays. We provided students with the following instructions for writing their individual design essays: "*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*". In introducing the activity, the teachers gave students general tips for writing clear and concise essays and information about the kinds of science relationships they should include in their essays. Students wrote their initial essay and received personalized feedback from PyrEval about the science content in their essays. Teachers, again, provided prompts for general tips for revising science ideas as well as clarity in writing. Students then participated in a peer editing activity using sample feedback and essays and were then given time to revise their essay using this feedback.

We tracked the extent to which students revised their individual essays. For this paper, we created two categories for tracking revisions: 1) *made no or minimal revisions* - such as making superficial changes to spelling or grammar or the addition of a single word or short dependent clause to an existing sentence (e.g., addition of the word "total" to the clause "so the energy stays the same"), and 2) *made substantive revisions* - such as adding new, content-related sentences or independent clauses that could stand alone as their own unit (e.g., adding "Although the energy might change forms the total energy stays the same due to the Law of Conservation of Energy"). These more substantive revisions could be a new idea or update of an existing, or removal of ideas. We did not assess the quality or scientific accuracy of the revisions for this analysis. Two researchers independently examined 20% of students' initial and revised essays to determine whether students made substantive revisions (or not) and achieved a 90% agreement. These researchers then discussed all discrepancies, and all disagreements were resolved.

Another source of data derived from students' design essays were the CU vector scores produced by PyrEval. We used these to create a *total initial essay score* for the initial essay and a *total revised essay score* for the revised essay by summing the total number of CUs an essay received, with a maximum score of 6. How these CU scores and the revision codes were used are discussed in more detail in the Analyses section.

Physics prior knowledge content test

To assess students' physics knowledge just before the unit started, our research team developed a physics content knowledge test consisting of 11 multiple-choice questions. We designed the test to assess students' understanding of the relationships among physical science ideas relevant to the roller coaster unit, such as how height and mass impact energy, as well as ideas about the conservation and transformation of energy within systems. Students received one point for correct answers on 10 of the 11 questions and incorrect answers received a zero. However, for one of the questions, students' answers could be scored as 0.5 for partially correct, 1 for fully correct, and 0 for incorrect. The total maximum score students could receive on the test was 11 points.

Analyses

We conducted three analyses: i) a hierarchical linear modeling (HLM) to explore the factors that influenced students' perceptions of the AI feedback, ii) a logistic regression to understand whether students' perceptions influenced their revisions or not, and iii) a linear regression to evaluate the extent to which students' perceptions predicted their revised essay scores. Because some data sources functioned as independent and dependent variables in different analyses, we listed our various sources and their respective roles in the analyses in Table 2.

Table 2

Data sources and the role they play in our different analyses.

Source	Data	Analyses
Perceptions questionnaire	Perception score	- Outcome variable in HLM - Independent variable in logistic regression - Independent variable in linear regression
CU scores produced by PyrEval	Initial essay score (E1total)	- Independent variable in HLM - Covariate in logistic regression - Covariate in linear regression
	Revised essay score	- Outcome variable in logistic regression - Outcome variable in linear regression
Physics pretest	Pretest score	- Covariate in HLM
Students' revisions	Binary revision code: no or minimal vs. substantive revision	- Outcome variable in logistic regression - Independent variable in linear regression

Exploring factors influencing students' perceptions of AI feedback

To understand the factors influencing students' perceptions (RQ1), we conducted a two-level HLM analysis to account for the nested nature of the data, i.e., students (n_{L1}=339) as level-1, nested within different teachers (n_{L1}=6) as level-2. Thus, the hierarchical modeling accounts for students with the same teacher being subject to the same instruction and environment, which may not be adequately captured using a linear regression model (Snijders & Bosker, 2012). We used students' perception scores (L1 variable) as the outcome variable and students' pretest score, initial essay score, and gender as the L1 explanatory variables. We selected a random-intercept model, allowing the intercept to vary between the different teachers. To estimate the proportion of variance attributed to the teacher (L2) and student (L1) levels, we first fitted a null random-intercept model, i.e., a model without student-level (L1) predictors (Model 1). Model 2 expanded on Model 1 by adding L1 predictors. The Level 1 and 2 equations as well as the final mixed model equation for Model 2 are presented in Table 3. The models were estimated with R software (R Core Team, 2024) using the *lme4* package (Bates et al., 2015). We also calculated the intraclass correlation (ICC) to determine the proportion of total variance of students' perceptions that is accounted for by the teacher level. The ICC was 0.02, indicating that most of the variability was attributable to students' individual characteristics, instead of teachers.

Table 3

Equations for HLM models.

Models	Equations
Level 1	$\text{perception\_score}_{ij} = \beta_{0j} + \beta_{1j}(\text{E1\_score})_{ij} + \beta_{2j}(\text{pretest\_score})_{ij} + \beta_{3j}(\text{male})_{ij} + R_{ij}$
Level 2	$\beta_{0j} = \gamma_{00} + U_{0j}$
	$\beta_{1j} = \gamma_{10}$
	$\beta_{2j} = \gamma_{20}$
	$\beta_{3j} = \gamma_{30}$
Mixed	$\text{perception\_score}_{ij} = \gamma_{00} + \gamma_{10}\,\text{E1 score}_{ij} + \beta_{2j}\text{pretest\_score}_{ij} + \beta_{3j}\text{male}_{ij} + R_{ij} + U_{0j}$

Exploring how students' perceptions may have influenced revisions

We were also interested in examining whether students' perceptions about the AWF impacted their revisions and writing improvement. To determine whether students with different perceptions engaged in making substantive revisions or not, we conducted a logistic regression, with substantive revision (or not) as the binary outcome variable, and students' perception scores and initial essay scores as explanatory variables (see Table 2). To understand whether students' perceptions of the automated feedback influenced their growth in writing, we fitted a linear regression model with students' revised essay scores as the outcome variable, the binary revise code and perception scores as the independent variables, and their initial essay score as the covariate.

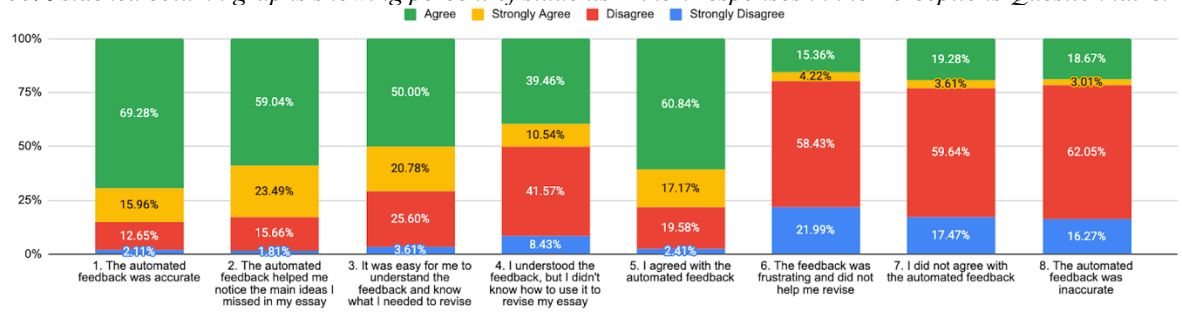
Results

Descriptive overview of students' perceptions, revisions, and CU scores

We found that most students had positive perceptions about the automated feedback in helping them revise. Out of the 339 students for whom we had full data, about 22.7% were strongly positive (scoring 26-32 points), 64.6% held at least mostly positive perceptions (scoring 20-25 points), 11.8% held mostly negative perceptions (scoring 14-19 points), and less than 1% held strongly negative views (scoring 8-13 points). To provide deeper insights into which aspects of the automated feedback students were positive or negative about, we also provide an overview of student responses to each Likert statement (Figure 2). We observed that students mostly agreed or strongly agreed with positively stated items (i.e., questions 1, 2, 3, and 5) and mostly disagreed or strongly disagreed with negatively stated items (i.e., questions 6, 7, and 8). About half the students said they understood the feedback, but were not sure how to use it to revise (question 4).

Figure 2

100% stacked column graphs showing percent of students' Likert responses in the Perceptions Questionnaire.



More students engaged in substantive revisions of science content (78.5%) than making no or minimal revisions (21.5%). We also found that students included significantly more CUs in their revised (*M=5.01, SD=1.09*) versus initial essays (*M=4.45, SD=1.33*); (t(338)=-11.888, p<0.001), using a paired samples t-test.

Factors influencing students' perceptions of AI feedback

In exploring the factors that may have influenced students' perceptions of the automated feedback (RQ1), we found that adding fixed effects (Model 2) to our HLM analysis significantly improved the random intercept model (Model 1) fit (see Table 4). We observed that some variability in students' perception scores was attributable to teachers (γ_{00} = 0.226), but most of the variation was at the student level (σ^2 = 11.484). Our results indicated that students' initial essay scores (*E1Total*) had a significant positive relationship with their perception scores (γ_{10}=0.367, *p*=0.0141), suggesting that a one-unit increase in initial essay scores was associated with a 0.367-unit increase in perception scores. We found no other significant effects on perception scores in our final model (Model 2). Based on the results in Table 4, we rewrote the equation as: $perception\_score_{ij} = 0.226 + 0.367\,E1\_score_{ij} + 0.095\,pretest\_score_{ij} + 0.456\,male_{ij} + R_{ij} + U_{0j}$.

Table 4

HLM results.

Fixed effect	Model 1 Coefficient (SD)	Model 2 Coefficient (SE)	*t* (df)	p
Intercept (γ_{00})	-	20.786(0.799)	26.025 (143.5)	<0.0000***
E1total (γ_{10})	-	0.367 (0.149)	2.469 (317.9)	0.0141*
pretest_score (γ_{20})	-	0.095 (0.102)	0.934 (315.3)	0.3508
male (γ_{30})	-	0.456 (0.388)	1.176 (317.9)	0.2406
Residual (σ^2)	12.459 (3.529)	11.484 (3.389)	-	-
Intercept (τ^2)	0.236 (0.485)	0.226 (0.475)	-	-

(Note: E1total range: 0-6; pretest range: 0-11)

Influence of students' perceptions on revisions and growth in writing

Our logistic regression analysis examined whether students' perceptions of automated feedback predicted the odds of making substantive revisions (RQ2). Our model met all assumptions for conducting a logistic regression. We found that the likelihood ratio test was significant ($\chi^2(2)=6.76,\ p<.05$). In this model, while students'

perception scores did not significantly predict whether they made substantive revisions or not (β= *-0.014, SE = 0.038, z = -0.361, p = .718*), we found that the higher a student's initial essay score was, they were significantly less likely to make a substantive revision (β= -0.271, SE = 0.112, z = -2.426, p = .015). Thus, for every one-point increase in initial essay score, the odds of making a substantive revision significantly decreased by 23.7% (*exp(-0.271)*). Based on the results, the logistic regression model equation was: *logit(revise) = 2.855 – (0.014 * perception score) – (0.271 * initial essay score)*.

We then conducted a multiple linear regression analysis to evaluate the extent to which students' perceptions could predict their growth in writing as assessed by their revised essay scores, when controlling for their initial essay scores and whether they revised or not (RQ2). We found that that all model assumptions were met except for linearity and homoscedasticity, which may be explained by the binary nature of the *revise variable*. As we found no significant outliers, we continued our analysis with this model. We found that students with higher perception scores made significantly greater gains in their revised essay scores than students with less positive perceptions (R^2=.6013, $F_{(3,335)}$=168.4, p<.01). The R^2 of 0.60 indicated that students' perceptions of the automated feedback and their initial essay scores explained approximately 60% of the variance in their revised essay scores. Based on the results, the linear regression equation was: *Revised essay score = 1.322 + (0.025 * perception score) + (0.405 * revise) + (0.627 * initial essay score)*. In this model, when controlling for their initial essay scores, a one-unit increase in a student's perception score was significantly associated with a 0.025-unit increase in their revised essay score. Further, students who made substantive revisions, on average, scored 0.405 points higher in their revised essay than students who did not.

Discussion

One of the primary goals of providing students with automated feedback on their writing is to give them opportunities to receive timely, formative feedback that they can use to engage in thoughtful revisions of their writing over multiple drafts. However, students' willingness to engage in substantive revisions that lead to growth in their writing may be impacted by several factors. Like other studies (Wilson et al., 2021b; 2024), we found that most students held positive perceptions of the automated feedback they received from PyrEval on their science writing. Like Roscoe et al.'s (2017, 2018) findings, the majority of students in our study made substantive revisions no matter their perceptions. However, the most important finding of this paper is in line with Nunes et al.'s (2022) review; we found that students with more positive perceptions of the automated feedback made revisions that resulted in significant improvements in their writing (as assessed by PyrEval). We are stating this first to frame the rest of our discussion because while it is interesting to understand the factors that impact students' perceptions so we might be able to foster positive ones in students, it only matters if students' perceptions actually impact their revisions and growth in writing.

Aside from our findings that students' perceptions influenced their growth in writing, our results also point to the importance of the initial feedback students receive. Two of our findings indicated that the immediate feedback students received from PyrEval significantly impacted their perceptions of the usefulness of the feedback and the extent to which they engaged in revising. First, similar to Zhu et al. (2017; 2020), we found that students who received feedback indicating they included more CUs in their writing had significantly higher perceptions of the feedback, whereas prior knowledge, gender, and teacher had little impact. This finding makes sense, since it seems like basic human nature to hold more positive perceptions of an AI system that shows you are doing well. Second, our results identified that students whose feedback indicated they included more key ideas in their initial essay (i.e., had higher initial essay scores) were significantly *less* likely to make substantive revisions. This finding also seems logical in that if a student received feedback indicating they included all, or most, of the key scientific ideas in their essays from PyrEval, they may be less motivated to engage in extensive revisions thinking they already accomplished the goals of the assignment. In both cases, it is essential that teachers help students understand that AI systems are fallible and that it is imperative they evaluate whether the AI is correct (or not) and reflect on how to improve their writing, no matter what the feedback indicates. Thus, feedback should be designed to motivate students who did well to continue improving, as well as providing struggling students with feedback to support and encourage them to use the feedback as an opportunity for growth, rather than developing negative perceptions of the process. This may mean including not only feedback about areas for improvement, but also metacognitive prompts outlining that the main objective in using automatic feedback is to support them in deeply engaging in the revision process to continually improve their writing by making it more concise, clear, and cohesive. This is especially important, because, as we found, students who engaged in substantive revisions on average scored 0.405 points higher (as assessed by PyrEval) in their revised essay than students who did not.

Limitations, implications, future research, and conclusions

As with all research studies, there are several limitations to consider when evaluating our findings. One limitation of this study is that the substantive versus no or minimal revision categories only captured macro differences between the extent to which students engaged in revision. Future work examining the types and depth of these revisions and whether the revisions addressed the feedback from the AWF system may reveal more about how students' perceptions may have influenced the kinds of revisions and improvements students made.

Another limitation has to do with the small number of teachers we had in our HLM model. It is more conventional to run HLM when you have a greater number of cases for a Level 2 variable than we had in this study. While we did not find that teacher was a significant predictor influencing students' perceptions, perhaps if we had a greater sample size, our findings would have been different. Relatedly, this paper did not provide information about how the teachers integrated the AWF system in their instruction. Teachers likely play an important role in how students use automated feedback systems (Delgado et al., 2024), as they play a critical role in meaningfully integrating technologies in classroom practices and routines and scaffolding students' use of it. Clearly, simply providing students with AWF is not enough. The goals of using these systems are not self-evident and students may not use them in intended ways if they have not been provided with knowledge about how they can utilize them to facilitate their learning and improvement. This seems to be especially true for students receiving more positive or more constructive feedback. For example, if less proficient writers or students who are having difficulties with their comprehension consistently receive feedback indicating that multiple revisions are necessary to fulfill expectations, they may develop negative perceptions of the system's feedback. This may lead them to disengage from participating meaningfully in the revision process. On the other hand, if students choose not to revise because they received highly positive initial feedback, they are missing opportunities for further growth and may be under the misapprehension that their ideas are correct and well communicated, when in fact the AI may have made errors. Students must be supported not only to understand where they can improve, but also help them use the AI formative feedback to reflect on how to improve (Dey et al., 2024), rather than using it to simply revise to get positive feedback or a "high score" (Lottridge et al., 2021; Moore & MacArthur, 2016). Thus, more work must be done to better understand how to successfully integrate AWF systems as part of a distributed scaffolding system (Puntambekar, 2022; Puntambekar & Kolodner, 2005) in the classroom, to facilitate students' understanding of the goals for using these systems to engage in revision for supporting their learning and development.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48.

Beiki, M. (2022). Review of writing-related theories. *Cultural Arts Research and Development*, *2*(1), 27-33.

Dey, I., Gnesdilow, D., Passonneau, R., & Puntambekar, S. (2024). Potential Pitfalls of False Positives. In *International Conference on Artificial Intelligence in Education* (pp. 469-476). Cham: Springer Nature Switzerland.

Chen, Z., Chen, W., Jia, J., & Le, H. (2022). Exploring AWE-supported writing process: An activity theory perspective.

Collins, P., Tate, T. P., & Warschauer, M. (2019). Technology as a lever for adolescent writing. *Policy Insights from the Behavioral and Brain Sciences*, *6*(2), 194-201.

Delgado, A., Wilson, J., Palermo, C., Cordero, T. M. C., Myers, M. C., Eacker, H. Potter, A., Coles, J., & Zhang, S. (2024). Relationships between middle-school teachers' perceptions and application of automated writing evaluation and student performance. In M.D. Shermis and J. Wilson (Eds.), *The Routledge international handbook of automated essay evaluation* (pp. 261-277). Routledge.

Fleckenstein, J., Liebenow, L. W., & Meyer, J. (2023). Automated feedback and writing: A multi-level meta-analysis of effects on students' performance. *Frontiers in Artificial Intelligence*, *6*, 1162454.

Fu, Q. K., Zou, D., Xie, H., & Cheng, G. (2024). A review of AWE feedback: Types, learning outcomes, and implications. *Computer Assisted Language Learning*, *37*(1-2), 179-221.

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).

Gnesdilow, D., Dey, I., Gengler, D., Malkin, L., Puntambekar, S., Passonneau, R., & Kim, C. (2024). The Impact of Middle School Students' Writing Quality on the Accuracy of the Automated Assessment of Science Content. In Lindgren, R., Asino, T. I., Kyza, E. A., Looi, C. K., Keifert, D. T., & Suárez, E. (Eds.), *Proceedings of the 18th International Conference of the Learning Sciences - ICLS 2024* (pp. 250-257). International Society of the Learning Sciences.

Graham, S. (2019). Changing how writing is taught. *Review of Research in Education*, *43*(1), 277-303.
Lottridge, S., Godek, B., Jafari, A., & Patel, M. (2021). Comparing the robustness of deep learning and classical automated scoring approaches to gaming strategies. Technical report, Cambium Assessment Inc.
Liu, Y., Xiong, W., Xiong, Y., & Wu, Y. F. B. (2024). Generating timely individualized feedback to support student learning of conceptual knowledge in Writing-To-Learn activities. *Journal of Computers in Education*, *11*(2), 367-399.
Moore, N. S., & MacArthur, C. A. (2016). Student use of automated essay evaluation technology during revision. *Journal of Writing Research*, *8*(1), 149-175.
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.
Ofosu-Ampong, K. (2023). Gender differences in perception of artificial intelligence-based tools. *Journal of Digital Art & Humanities*, *4*(2), 52-56.
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. (2022). Distributed scaffolding: Scaffolding students in classroom environments. *Educational Psychology Review*, *34*(1), 451-472.
Puntambekar, S., & Kolodner, J. L. (2005). Toward implementing distributed scaffolding: Helping students learn science from design. *Journal of Research in Science Teaching, 42*(2), 185–217.
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.
Roscoe, R. D., Allen, L. K., Johnson, A. C., & McNamara, D. S. (2018). Automated writing instruction and feedback: Instructional mode, attitudes, and revising. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 62, No. 1, pp. 2089-2093). Sage CA: Los Angeles, CA: SAGE Publications.
Singh, P., Passonneau, R. J., Wasih, M., Cang, X., Kim, C., & Puntambekar, S. (2022). Automated support to scaffold students' written explanations in science. In *International Conference on Artificial Intelligence in Education* (pp. 660-665). Cham: Springer International Publishing.
Snijders, T. A. B., & Bosker, R. J. (2012). Multilevel analysis: An introduction to basic and advanced multilevel modelling (2nd ed.). Sage.
Wilson, J., Ahrendt, C., Fudge, E. A., Raiche, A., Beard, G., & MacArthur, C. (2021a). Elementary teachers' perceptions of automated feedback and automated scoring: Transforming the teaching and learning of writing using automated writing evaluation. *Computers & Education*, *168*, 104208.
Wilson, J., Huang, Y., Palermo, C., Beard, G., & MacArthur, C. A. (2021b). Automated feedback and automated scoring in the elementary grades: Usage, attitudes, and associations with writing outcomes in a districtwide implementation of MI Write. *International Journal of Artificial Intelligence in Education, 31*(2), 234-276.
Wilson, J., Zhang, F., Palermo, C., Cordero, T. C., Myers, M. C., Eacker, H., Potter, A., & Coles, J. (2024). Predictors of middle school students' perceptions of automated writing evaluation. *Computers & Education (211)*104985.
Zhai, N., & Ma, X. (2023). The effectiveness of automated writing evaluation on writing quality: A meta-analysis. *Journal of Educational Computing Research*, *61*(4), 875-900.
Zhang, Z. V. (2020). Engaging with automated writing evaluation (AWE) feedback on L2 writing: Student perceptions and revisions. *Assessing Writing*, *43*, 100439.
Zhu, M., Lee, H. S., Wang, T., Liu, O. L., Belur, V., & Pallant, A. (2017). Investigating the impact of automated feedback on students' scientific argumentation. *International Journal of Science Education*, *39*(12), 1648-1668.
Zhu, M., Liu, O. L., & Lee, H. S. (2020). The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Computers & Education*, *143*, 103668.

Acknowledgements

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).