Potential Pitfalls of False Positives

Indrani Dey[1], Dana Gnesdilow[1], Rebecca Passonneau[2], and Sadhana Puntambekar[1]

[1] University of Wisconsin-Madison, Madison WI 53706, USA
{idey2,gnesdilow}@wisc.edu, puntambekar@education.wisc.edu
[2]The Pennsylvania State University, State College 16801, USA
rjp49@psu.edu

Abstract. Automated writing evaluation (AWE) systems automatically assess and provide students with feedback on their writing. Despite learning benefits, students may not effectively interpret and utilize AI-generated feedback, thereby not maximizing their learning outcomes. A closely related issue is the accuracy of the systems, that students may not understand, are not perfect. Our study investigates whether students differentially addressed false positive and false negative AI-generated feedback errors on their science essays. We found that students addressed nearly all the false negative feedback; however, they addressed less than one-fourth of the false positive feedback. The odds of addressing a false positive feedback was 99% lower than addressing a false negative feedback, representing significant missed opportunities for revision and learning. We discuss the implications of these findings in the context of students' learning.

Keywords: AI Accuracy, Automated Feedback, Science Writing.

1 Introduction

Studies on artificial intelligence (AI) systems are increasingly exploring how to support classroom activities, by automating routine parts of teachers' work, and allowing them to provide more meaningful support to students [1, 2]. However, AI systems are often imperfect, resulting in frustration and trust issues among users [3], possibly from an incomplete understanding of how AI works and its capabilities. To thrive in a world increasingly permeated by AI, which includes educational spaces, students need to develop a critical understanding of what AI is and how it works so they can competently interact with it and utilize AI-generated output [4]. While it is important that the output is accurate for students to use it effectively, AI systems are not infallible. Therefore, we need to better understand how students respond to errors in AI-generated output, to inform best practices for scaffolding the use of AI in educational contexts. This study investigates the extent to which students addressed false positive and false negative errors in AI-generated feedback in the context of revising their science writing.

AI technologies, particularly Natural Language Processing (NLP) techniques are increasingly being used to assess students' writing, particularly essays that are hard to assess in a timely manner [5], such as written explanations of scientific phenomena that students often struggle with. This automated feedback, tailored to individual student needs, allows for ongoing formative assessment [6]. Further, teachers can use the AI-generated feedback to identify gaps in students' understanding and drive instruction, while students can use the feedback to critically evaluate and refine their work, thereby deepening their understanding and improving their learning outcomes.

The accuracy of automated feedback on writing is an ongoing challenge. Building realistic expectations and understanding of AI-generated outputs in end-users plays a role in their user experience and acceptance of AI technologies [7], which, in turn, can impact their engagement and learning. False positive errors, where an incorrect answer is marked as correct, can be particularly challenging to address, as students may not be able to identify the errors. Studies have found that students often ignore feedback from a false positive output as they may not have identified it as missing or incorrect [8], and may be reluctant to correct something they thought they already had right [9], thus, missing opportunities to revise and learn. On the other hand, while false negative errors (where correct answers are marked as wrong) can lead to more engagement as students spend more time reviewing their answers [9], these errors more readily contribute to students' dissatisfaction and perceptions of unfairness [10], leading to trust issues with the system and a lower propensity for future tool use [3, 11].

While studies have explored students' perceptions and/or revision behaviors using automated feedback on short answers [9, 10] or writing quality on English essays [3, 8, 11], few have investigated student revision behaviors using automated feedback on the science content in essays; fewer have focused on the extent to which students address AI assessment and feedback errors when revising scientific essays. Our previous study found that although students included significantly more science ideas in their revised essays, the NLP system we used made errors about 25% of the time [12]. This study further investigates the nature of the errors –i.e., false positive versus false negative– and whether students addressed the feedback by revising ideas in their essays (or not) when they received erroneous feedback. The research questions guiding our study are:

1. What were the rates of false positive and false negative feedback errors generated by our NLP system?
2. When students received erroneous feedback, did they address the false positive or false negative feedback in similar ways?
3. To what extent did addressing erroneous feedback impact students' learning?

2 Methods

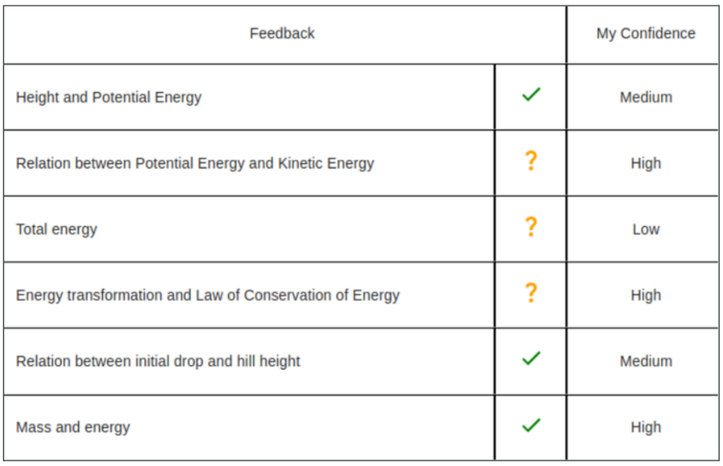
2.1 Study Design and Context

A total of 238 students from three 8th-grade public middle school science classrooms in the midwestern US participated in this study (n_1=96, n_2=80, and n_3=62). Students conducted experiments using a virtual roller coaster simulation to explore relationships between height, mass, and energy. They wrote an essay using data from their trials to explain the scientific phenomena behind their roller coaster design. They submitted these essays to our NLP system, PyrEval (described below), which assessed the essays and provided automated feedback. Students were supposed to use this feedback to identify missing ideas, revise, and submit their revised essays to PyrEval for re-assessment. Students were aware they were receiving AI-generated feedback that may not be 100% accurate. Students took the same multiple-choice test before and after the unit.

2.2 PyrEval Assessment and Feedback

PyrEval, the NLP system that provided automated feedback on students' writing in this study, uses a wise-crowd model to identify weighted vectors of key content ideas [13], known as content units or CUs; more important ideas are more highly weighted. For this unit, PyrEval identified 6 high-weighted CUs aligned with the important ideas or relationships students should have included in their essays (see Fig. 1).

Once students submitted their essays, PyrEval parsed each essay into separate sentences and examined each sentence for the presence or absence of each of the 6 highly weighted CUs. If PyrEval detected a certain CU, it would produce a vector score of 1, whereas a vector score of 0 indicated that PyrEval did not detect that CU in the essay. These vector scores were presented to the students in the form of a feedback chart (see Fig. 1), indicating which ideas may have been present or missing from their essays. The chart also provided a "My Confidence" column, indicating PyrEval's estimated accuracy in detecting or not detecting a particular CU in the essay. Students were asked to also attend to this column when considering what feedback to address when revising.



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

Fig. 1. Sample feedback chart, showing each CU, if it was detected (green check mark) or not (orange question mark) in the essay, and PyrEval's approximate accuracy about the detection.

The vector scores for each essay were recorded in the backend in the following (example) format: [1,0,0,1,0,1], with a 1 or 0 indicating the presence or absence of an idea in that essay, respectively. Thus, in the above example, PyrEval detected CUs 0, 3, and 5 in a particular essay, but did not detect CUs 1, 2, and 4. These vector scores were then used to assess student's performance for each essay as well as PyrEval's accuracy.

2.3 Data Sources and Analyses

Out of the students who submitted both an initial and revised essay, we randomly chose 20 students from each teacher's classes, for a total of 60 students and 120 corresponding essays (60 x (1 initial essay + 1 revised essay)). For an in-depth look into how students responded to erroneous automated feedback, we first determined the percent of false positive and false negative feedback errors in their initial essays and then examined whether students addressed the feedback by making revisions. As false positive errors can result in missed learning opportunities, we also used students' scores on the pre to post content knowledge test and their responses to erroneous feedback to understand how students' responses (or lack thereof) may have impacted their science learning.

PyrEval Accuracy. To assess PyrEval's accuracy, two researchers independently coded 20% of the 60 students' initial and revised essays for each content unit, and compared their codes to PyrEval's vector scores. If PyrEval determined the presence of a CU in an essay when it was absent, we considered it a *false positive error*. If PyrEval failed to detect a CU present in the essay, we considered it a *false negative error*. The two coders achieved substantial inter-rater reliability with Kappa = 0.768 [14]. Discrepancies were resolved through discussion and one researcher coded the remaining data. We calculated the percent error by adding the total false positive and false negative errors, dividing it by 360 (60 essays x 6 CUs per essay), and multiplying by 100.

Addressing Feedback. To assess whether students addressed the automated feedback or not, we parsed each student's initial and revised essay into separate sentences and then compared each sentence to examine the revisions. We then coded which CU was addressed for each revision, i.e., each update to an existing idea or an addition of a new idea. Two researchers independently coded 15% of the 60 students' initial and revised essays to determine which CUs students addressed in their revisions, achieving almost perfect inter-rater reliability (Kappa = 0.911). One researcher then coded the remaining data. We used this data in a Chi-square test of independence to explore whether there were differences in the proportion of students who addressed false positive versus false negative feedback. We also used this data to perform a logistic regression to determine whether the type of feedback errors and students' initial essay scores would predict whether they would address erroneous feedback or not.

Physics Content Knowledge Test. Students took a multiple-choice content knowledge test assessing their understanding of the relationships between the height of the initial drop and mass of the roller coaster car and the amount of energy and speed on the ride. It also assessed their understanding of energy transformation and conservation. Students could score a maximum score of 11 points on the test. We used this data to explore the extent to which addressing erroneous feedback may have impacted students' science content learning, conducting a multiple linear regression analysis.

3 Results

We will first provide the descriptive statistics of PyrEval's performance in detecting CUs in the initial essays. We will then present the extent to which students addressed the PyrEval feedback, followed by the results of the Chi-square and regression analyses.

PyrEval Accuracy. PyrEval searched for a total of 360 CUs in students' initial essays. We found that PyrEval was 74.7% accurate in correctly identifying whether the students included or did not include the CUs. On the other hand, we found that PyrEval made errors 25.2% of the time, with a total of 91 errors. Of these errors, 73.6% were false positive errors and 26.4% were false negative errors. This means that PyrEval was nearly three times as likely to make a false positive rather than a false negative error.

Nature of Students' Revisions in Response to Automated Feedback. Of our 60-student sample, 96.7% revised by updating one or more existing ideas or adding new ideas. Two students made no revisions and resubmitted their initial essays. PyrEval identified a total of 105 CUs that were missing from all initial essays (including 24 false negatives), of which 76.1% were addressed in revisions. This indicated that the majority of students revised based on PyrEval's feedback that a CU was missing from their essays. However, there were 67 instances where PyrEval gave students false positive feedback. In this case, we found that only 15 of these false positive errors were addressed.

To examine whether students were significantly more likely to address false negative or false positive feedback errors using a chi-squared analysis of independence, we created a contingency table (Table 1) for the 91 total errors by PyrEval, indicating whether errors were false positive or false negative and if students addressed them or not.

Table 1. Contingency table showing the frequency of whether students addressed or did not address false positive or negative errors made by PyrEval.

	Addressed feedback	Did not address feedback	Total
False positive	15 (*0.22*)	52 (*0.78*)	67
False negative	22 (*0.92*)	2 (*0.08*)	24
Total	37	54	91

Table 1 shows that while 92% of the students addressed false negative feedback, only 22% addressed false positive feedback. We found that a significantly higher proportion of students addressed false negative feedback than false positive feedback ($X^2_{(1, N=91)} = 35.150, p < 0.0001$) when revising their essays.

To ensure these findings were not based on students' performance on their initial essays or the teacher they had, we performed a logistic regression. In the model, the likelihood of a feedback being addressed or not was considered the outcome variable, the error type as the independent variable, and students' initial essay CU score and their teacher as the covariate. The logistic regression output revealed a significant coefficient for addressing false positive feedback (-3.62233), when addressing false negative feedback was the baseline category (*$p<0.05$*). The odds of a student addressing a false positive feedback was 99% lower than addressing a false negative feedback, if all other variables are constant, given an odds ratio of exp(-3.62233) = 0.026.

Thus, students were overwhelmingly more likely to address a false negative error than a false positive error, leading us to investigate whether there was a relationship between failing to address a false positive error and students' science learning. We focused only on false positive feedback errors for two reasons: first, students addressed most of the false negative feedback (see Table 1); second, studies found that failing to address false positive feedback on students' short answers negatively impacted learning outcomes [9], and we wanted to explore this for students' longer writing pieces.

Relationship between not addressing false positive feedback and science learning. Of the 60 students in our sample, 43 received at least one false positive feedback. Of these forty-three students, 34 had taken both the pre and post-tests. Thus, we had the complete data for these 34 students, which was then used for the multiple linear regression. We calculated the percentage of false positive feedback students did not address by dividing the total number of false positives each student addressed by the total number of false positives received, then multiplied by 100.

We conducted a multiple linear regression, using students' post-test scores as the dependent variable, the percent of false positive errors addressed as the independent variable, and their pre-test score and initial essay score as covariates. The model explained 17.9% of the variance ($R^2=0.1786$, $F_{(3,30)}=2.175$, $p=0.1$). While there was a negative relationship between the percent of false positive errors not addressed and post-test scores, it was not statistically significant ($x=-0.5844$, $p=0.3415$). The model also indicated slight positive relationships with pre-test scores ($x= 0.1783$, $p=0.1581$) and essay scores ($x=0.3659$, $p=0.0692$), but neither were statistically significant.

4 Discussion

While automated writing assessments can support students' learning, students may not effectively interpret and utilize the AI-generated feedback, especially when the system provides inaccurate feedback. Our study investigated whether students responded differently in addressing AI-generated feedback with false positive or false negative errors, to understand if students were critically examining the AI feedback to make targeted revisions based on what was truly missing in their essays.

We found that PyrEval made errors about the presence or absence of science ideas in essays about one-fourth of the time and that it was three times more for a false positive error than a false negative error. Not only was the potential for receiving a false positive error much higher than a false negative error, students were significantly more likely to address a false negative error than a false positive error, thus magnifying the potential for missed learning opportunities.

The majority of students had addressed at least some of PyrEval's feedback, suggesting that they were not averse to making revisions, as also seen in other studies [3, 11]. Therefore, false positive feedback errors represent lost opportunities for students to address missing ideas. Although we found that students who did not address false positive feedback had lower learning outcomes, unlike other studies, it was not significant [9]. This may be due to a few reasons. First, there may be differences in students' revisions on shorter writing assignments [9] versus long essays and in different subjects. Second, our sample size was limited. Third, since the pre and post-tests were identical, recall bias may have further affected the internal validity. Another potential confound could be that students participated in other science activities apart from writing their essays in this unit, which may have influenced their conceptual learning. Further, students may differ in their understanding of latent relationships between CUs, the examination of which was beyond the scope of this study (e.g., students do not mention CU2 if they mention CU3). Future research may address and expand on these issues.

Although students addressed all false negative errors in our study, it could create trust issues with the feedback. This may lead to disengagement with the AI systems integrated with schoolwork, which in turn, could affect learning outcomes. Other studies investigating user experiences with AI systems also found differences in users' trust, participation with, or acceptance of the AI system, such as a conversational agent [15] or a scheduling assistant [7]. These studies highlight imperfections in AI outputs and discuss the trade-off between favoring a false positive versus negative error. Despite reducing the error in potentially more harmful contexts, there is still a reliance on technology that may cause other unanticipated issues in students' participation or learning.

Thus, helping learners understand how AI output may contain potential errors and how to identify and address them, may help mitigate some of the effects of both false positive and false negative errors, as well as provide students with more agency over their learning [4]. Instead of passively following automated feedback, students can be encouraged to work in partnership with the technology to mindfully engage with the feedback and critically apply it to improve and develop understanding [16]. Thus we recommend that developing these competencies should be considered as a part of AI-literacy as well as AI-related curricula. Our future work will be to investigate students' perceptions and actions on using AI-generated output and help inform how to help learners more effectively use automated feedback to maximize their learning.

Acknowledgments. We thank the students and teachers who participated in this study.

References

1. Holstein, K., McLaren, B. M., Aleven, V.: Co-designing a real-time classroom orchestration tool to support teacher-AI complementarity. Journal of Learning Analytics 6, 27–52 (2019a).
2. Holstein, K., McLaren, B. M., Aleven, V.: Designing for complementarity: Teacher and student needs for orchestration support in AI-enhanced classrooms. In Artificial Intelligence in Education: 20th International Conference, AIED 2019, vol. 20, pp. 157–171. Springer International, (2019).
3. Roscoe, R. D., Snow, E. L., McNamara, D. S.: Feedback and revising in an intelligent tutoring system for writing strategies. Artificial Intelligence in Education: 16th International Conference, AIED 2013, vol. 16, pp. 259–268. Springer, Heidelberg (2013).

4. Markauskaite, L., Marrone, R., Poquet, O., Knight, S., Martinez-Maldonado, R., Howard, S., Tondeur, J., De Laat, M., Buckingham Shum, S., Gašević, D., Siemens, G.: Rethinking the entwinement between artificial intelligence and human learning: What capabilities do learners need for a world with AI? Computers and Education: Artificial Intelligence 3, 100056 (2022).
5. Ramesh, D., Sanampudi, S. K.: An automated essay scoring systems: a systematic literature review. Artificial Intelligence Review 55(3), 2495–2527 (2022).
6. Ai, H.: Providing graduated corrective feedback in an intelligent computer-assisted language learning environment. ReCALL, 29(3), 313–334 (2017).
7. Kocielnik, R., Amershi, S., Bennett, P. N.: Will you accept an imperfect AI? Exploring designs for adjusting end-user expectations of AI systems. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–14 (2019).
8. Dodigovic, M., Tovmasyan, A.: Automated writing evaluation: The accuracy of Grammarly's feedback on form. International Journal of TESOL Studies 3(2), 71–87 (2021).
9. Li, T. W., Hsu, S., Fowler, M., Zhang, Z., Zilles, C., Karahalios, K.: Am I Wrong, or Is the Autograder Wrong? Effects of AI Grading Mistakes on Learning. In: Proceedings of the 2023 ACM Conference on International Computing Education Research, vol. 1, pp. 159–176 (2023).
10. Hsu, S., Li, T. W., Zhang, Z., Fowler, M., Zilles, C., Karahalios, K.: Attitudes surrounding an imperfect AI autograder. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1–15 (2021).
11. Roscoe, R. D., Wilson, J., Johnson, A. C., Mayra, C. R.: Presentation, expectations, and experience: Sources of student perceptions of automated writing evaluation. Computers in Human Behavior 70, 207221 (2017).
12. Gnesdilow, D., Dey, I., Gengler, D., Malkin, L., Puntambekar, S., Passonneau, R.J., & Kim, C.: The impact of middle school students' writing quality on the accuracy of the automated assessment of science content. In: Proceedings of ISLS (2024).
13. Gao, Y., Chen, S., Passonneau, R. J.: Automated Pyramid Summarization Evaluation. In: Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), pp. 404–418. Association for Computational Linguistics (2019).
14. Stemler, S.: An overview of content analysis. Practical Assessment, Research & Evaluation 7(17), 137–146 (2001).
15. Do, H. J., Kong, H. K., Tetali, P., Lee, J., Bailey, B. P.: To Err is AI: Imperfect Interventions and Repair in a Conversational Agent Facilitating Group Chat Discussions. In: Proceedings of the ACM on Human-Computer Interaction 7, pp. 1–23 (2023).
16. Salomon, G., Perkins, D. N., Globerson, T.: Partners in cognition: Extending human intelligence with intelligent technologies. Educational Researcher 20(3), 2–9 (1991).