International Society of the Learning Sciences

NLP-Enabled Automated Feedback about Science Writing

ChanMin Kim, Pennsylvania State University, cmk604@psu.edu
Sadhana Puntambekar, University of Wisconsin – Madison, puntambekar@education.wisc.edu
Eunseo Lee, Pennsylvania State University, eul374@psu.edu
Dana Gnesdilow, University of Wisconsin – Madison, gnesdilow@wisc.edu
Mahsa Sheikhi Karizaki, Pennsylvania State University, mfs6614@psu.edu
Rebecca J. Passonneau, Pennsylvania State University, rjp49@psu.edu

Abstract: Eighth grade students received automated feedback from PyrEval - an NLP tool - about their science essays. We examined essay quality change when revised. Regardless of prior physics knowledge, essay quality improved. Grounded in literature on AI explainability and trust in automated feedback, we also examined which PyrEval explanation predicted essay quality change. Essay quality improvement was predicted by high- and medium-accuracy feedback.

Introduction

Teaching students to engage in scientific practices such as writing is core to the Next Generation Science Standards, but is by no means easy (Osborne, 2014). To do so requires both adequate support, and actionable feedback. Automated feedback can provide customized feedback on students' scientific explanations (e.g., Gerard & Linn, 2022). However, little is known about the role of prior science knowledge in using automated feedback. While personalized feedback specially designed for students with low prior knowledge was used in the context of using automated feedback (Tansomboon et al., 2017), methods of designing automated feedback in discrete and effective manners for students who begin with low science knowledge are needed. Still, students do not always use automated feedback to inform revisions of their work, especially when they do not trust automated feedback (Conijn et al., 2023; Ranalli, 2021). Lower-performing students often ignore feedback, potentially due to low expectancies for improvement (Tansomboon et al., 2017). To enhance trust in automated feedback, transparency about automated feedback's approach and accuracy is recommended (Ranalli, 2021; Tansomboon et al., 2017).

The conceptual framework of this study was informed by literature on automated assessment of scientific explanations (e.g., Gerard & Linn, 2022), and explainability of AI and user trust (e.g., Conijn et al., 2023; Ranalli, 2021). Research questions were: Does essay quality change when revised? Does essay quality change vary according to prior physics knowledge? How does PyrEval accuracy explanation predict essay quality change?

Method

Seven 8th grade science teachers from two US school districts and their students participated in a design-based physics unit for 14 to 15 instructional periods (45-min. each). Students wrote essays about relationships between height, mass, and energy, energy transformation, and the law of conservation of energy while designing a roller coaster, and revised their essays using automated feedback generated by PyrEval (Singh et al., 2022). PyrEval assessed the extent to which students addressed six content units (**CU**s). PyrEval assigned a score of 1 if present or 0 if absent for each CU. The scores were used to list either a check mark (✓) or a question mark (?) in a feedback table (Figure 1). We used a training set of essays to test PyrEval's accuracy, resulting in a set of labels regarding PyrEval's accuracy (high, medium, low) for each CU, which were listed in the feedback table.

We analyzed 337 students' essay quality scores calculated by summing up PyrEval-generated scores for each CU. Prior physics knowledge scores were grouped into low (29%, $M = 3.11$, $SD = 0.96$), moderate (45%, $M = 5.50$, $SD = 0.72$), and high levels (26%, $M = 7.76$, $SD = 0.87$) based on pretest scores (possible scores ranged from 0 to 11). The pretest scores were normally distributed (Shapiro-Wilk's $p = 0.058$).

Results and discussion

Repeated measures ANOVA indicated that there was a significant increase in essay quality between version 1 ($M = 4.35$, $SD = 1.42$) and version 2 ($M = 5.10$, $SD = 1.18$), $F(1, 334) = 169.323$, $p < 0.001$, Cohen's $d = 0.71$. There was also a significant difference in essay quality among the low ($M = 4.46$, $SD = 2.26$), moderate ($M = 4.73$, $SD = 1.78$), and high ($M = 4.98$, $SD = 2.24$) prior physics knowledge groups, $F(2, 334) = 4.614$, $p = 0.011$, $\eta_p^2 = 0.027$. The high prior knowledge group demonstrated significantly higher essay quality than the low prior knowledge group, $p < 0.01$. However, there was no significant interaction between time and prior knowledge level, $p = 0.169$. Regardless of prior physics knowledge, essay quality improved when revised (Figure 2).

Figure 1

Sample Essay (left) and Automated Feedback Table (right)

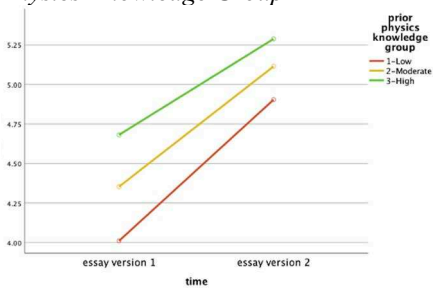
We started at a release height of 2 then we tested 3,4 and 5. each time we tested the roller coaster and **the greater the height was greater the PE at the top**. But the KE at the top was always 0. **As the cart went down the hill the PE went down and the KE went up. The energy basically switched** so if the PE at the top was 997 J the KE at the top was 0 J then the PE at the bottom was 0 J and the KE at the bottom was 977 J. **The total energy is always all the energy added up and PE is energy before the cart goes down the hill and KE is energy after the hill.** When we added the hill we learned that **the hill always had to be smaller then the first drop** or the cart would not make it over the hill. The weight of the cart has a lot to do with if the cart makes it to the finish or not **if the cart was heavier it had more energy.**

Note: CU0 in green text CU1 text in blue text CU2 in red text CU3 in purple text CU4 in pink text CU5 in orange text

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

Figure 2

Essay Quality Change per Prior Physics Knowledge Group



To examine which explanation about PyrEval predicts essay quality change, we grouped explanations into high, medium, and low accuracy explanations (Figure 1), and ran a GLMM with the model: essay quality change ~ time + prior physics knowledge + CU score change with high accuracy explanation + CU score change with medium accuracy explanation + CU score change with low accuracy explanation + (student). The results showed that score changes in the CUs within high- and medium-accuracy explanations were significant predictors of essay quality change, Ps < 0.001, but not in the CU with low-accuracy explanation, $p = 0.136$.

While further research is warranted, these findings suggest that automated feedback was effective for all three prior knowledge groups. Simple visualization and applying asset-based approaches may have facilitated engagement of students with low prior knowledge in revision. For example, when students saw a question mark, it did not mean the absence of the CU to them; rather, it meant that they needed to analyze their own essay to see if the CU was indeed missed or PyrEval was inaccurate. This approach may have steered students away from seeing deficit in their ability and motivated them to revisit their writing. Still, students would likely discard feedback that they do not trust (Conijn et al., 2023; Ranalli, 2021). Considering recent studies in which simple visualization increased user understanding and trust in AI (e.g., Branley-Bell et al., 2020; Leichtmann et al., 2023), explanations about PyrEval's accuracy through simple visualization may have helped with students' feedback understanding and trust. In essence, actionable information was provided to students that they could use to calibrate their trust appropriately and inform their use or nonuse of the feedback (Ranalli, 2021).

References

Branley-Bell, D., Whitworth, R., & Coventry, L. (2020). User trust and understanding of explainable AI: Exploring algorithm visualisations and user biases. In M. Kurosu (Ed.), *Human-Computer Interaction. Human Values and Quality of Life* (pp. 382–399). Springer International Publishing.

Conijn, R., Kahr, P., & Snijders, C. (2023). The effects of explanations in automated essay scoring systems on student trust and motivation. *Journal of Learning Analytics*, *10*(1), 37–53.

Gerard, L., & Linn, M. C. (2022). Computer-based guidance to support students' revision of their science explanations. *Computers & Education*, *176*, 104351. https://doi.org/10.1016/j.compedu.2021.104351

Leichtmann, B., Humer, C., Hinterreiter, A., Streit, M., & Mara, M. (2023). Effects of explainable artificial intelligence on trust and human behavior in a high-risk decision task. *Computers in Human Behavior*, *139*, 107539. https://doi.org/10.1016/j.chb.2022.107539

Osborne, J. (2014). Teaching scientific practices: Meeting the challenge of change. *Journal of Science Teacher Education*, *25*(2), 177–196. https://doi.org/10.1007/s10972-014-9384-1

Ranalli, J. (2021). L2 student engagement with automated feedback on writing: Potential for learning and issues of trust. *Journal of Second Language Writing*, *52*, 100816. https://doi.org/10.1016/j.jslw.2021.100816

Singh, P., Passonneau, R. J., Wasih, M., Cang, X., Kim, C., & Puntambekar, S. (2022). Automated support to scaffold students' written explanations in science. In M. M. Rodrigo, N. Matsuda, A. I. Cristea, & V. Dimitrova (Eds.), *Artificial Intelligence in Education* (pp. 660–665). Springer International Publishing.

Tansomboon, C., Gerard, L. F., Vitale, J. M., & Linn, M. C. (2017). Designing automated guidance to promote productive revision of science explanations. *International Journal of Artificial Intelligence in Education*, *27*(4), 729–757. https://doi.org/10.1007/s40593-017-0145-0

Acknowledgments

This work was supported by Grants 2010351 and 2010483 from the National Science Foundation (USA).

