International Society of the Learning Sciences

Examining the Effect of Automated Assessments and Feedback on Students' Written Science Explanations

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

Abstract: Writing scientific explanations is a core practice in science. However, students find it difficult to write coherent scientific explanations. Additionally, teachers find it challenging to provide real-time feedback on students' essays. In this study, we discuss how PyrEval, an NLP technology, was used to automatically assess students' essays and provide feedback. We found that students explained more key ideas in their essays after the automated assessment and feedback. However, there were issues with the automated assessments as well as students' understanding of the feedback and revising their essays.

Introduction

Engaging students in authentic science practices is central to learning and understanding science (NGSS, 2013; Braaten & Windschitl, 2011). At the core of such practices is writing scientific explanations, which researchers say should: (1) use data and/or evidence to make claims, and (2) connect scientific principles to evidence to explain the observed phenomenon (Berland & Hammer, 2012; Berland et al., 2016; Krajcik et al., 2014). Natural Language Processing (NLP) technologies can support students in writing science explanations by providing automated feedback (Gerard & Linn, 2016). This study investigated the role of an NLP technology, PyrEval, (Gao et al., 2018; Passonneau et al., 2018), to assess students' written science explanations. The research questions guiding our study were:

1. How does automated feedback help students explain key ideas in their science essays?
2. What are the opportunities and limitations of using automated feedback in classrooms?

Methods

Participants and context

A total of 264 students from three 8th-grade public middle school classrooms in the Midwestern US participated in this study. Students learned the science of roller coasters by conducting experiments in a simulation and writing essays to develop a roller coaster design based on the science they were learning. Students wrote a design essay (E1) after conducting their first set of experiments, which was sent to PyrEval for feedback. Next, students conducted additional experiments and wrote a second design essay (E2) that built on ideas and feedback in E1. They received feedback from PyrEval on their E2, based on which, they were given the opportunity to make revisions and resubmit their final design essays. We called this essay 2 revised (E2R).

PyrEval uses a wise-crowd model, where samples used to match key ideas, referred to as content units (CUs), are taken from a range of student essays. It automatically parses students' essays into propositions and creates a model of important propositions derived from a small set of reference responses. It then creates a vector representation for propositions and compares them to recognize paraphrases of similar content. In our study, 15 CUs were identified as the most important ideas for students to learn during the unit, which PyrEval applied to automatically assess students' essays.

Results and Discussion

We used the number of CUs identified by PyrEval in student's essays as the measure for our analysis. We conducted a repeated measures analysis to examine changes in students' essays from E1, to E2, to E2R after receiving feedback for students who completed all three essays ($N=228$). We summed the 15 CUs into a CU total score for each student's essays and used these as dependent outcomes. Our analysis showed that students included significantly more CUs in E2R ($M=5.05$, $SD=3.02$) than E2 ($M=4.77$, $SD=2.93$), and significantly more CUs in E2 than in E1 ($M=3.68$, $SD=2.40$) ($F_{1,228}=82.1889$; $p<.001$; $\eta p^2=.267$). We also ran Wilcoxon signed-rank tests to understand the changes in students' essays for each CU from Essay 1 to Essay 2 Revised. We found significant

differences between E1 and E2R for 6 of the CUs; this means more students included these CUs in E2R, compared to their E1. Conversely, there were no significant differences for the remaining CUs.

Our analysis showed that students included significantly more CUs from both E1 to E2 and E2 to E2R. This indicates that students included more ideas in their essays after receiving feedback, a finding also observed in other studies (Gerard et al., 2019; Tansomboon et al., 2017; Zhu et al., 2020). Our study provides evidence that PyrEval was able to expedite the 'scoring' of students' essays, which otherwise would have taken a long time to do manually for teachers, as has been observed in other studies evaluating NLP technologies to automatically assess students' science writing. However, there were some challenges: *First*, there were some challenges with PyrEval correctly identifying CUs. PyrEval looks for ideas, or CUs, in individual sentences. When students' ideas were scattered across multiple sentences or even paragraphs, the technology may not have been able to recognize the CUs. Students often write long sentences to explain and repeat ideas over several sentences, which often lack precision. They also sometimes forget to include punctuation. We found that PyrEval had difficulty recognizing ideas in sentences longer than 25 words. *Second*, we found that it is challenging to translate automatic assessments into comprehensible feedback for students.

Clearly, writing and revising explanations is not an easy task for middle school students. Using the findings from this study, we plan to make changes to our approach to ensure that we provide adequate scaffolding to students. We plan to improve PyrEval's ability to recognize CUs in students' writing by refining the NLP model, based on this year's classroom dataset. Additionally, our plan is to provide teachers with more information about students' progress through a teacher dashboard by summarizing trends about students' progress at multiple points in the unit, prior to and after essay writing. This high-level overview will allow teachers to see trends across their classes, enabling them to dynamically make titrated instructional decisions to ensure students get the support they need when they need it.

References

Braaten, M., & Windschitl, M. (2011). Working toward a stronger conceptualization of scientific explanation for science education. *Science Education*, *95*, 639–669.

Berland, L. K., Hammer, D. (2012). Framing for scientific argumentation. *Journal of Research in Science Teaching*, *49*, 68–94.

Berland, L. K., Schwarz, C., V., Krist, C., Kenyon, L., Lo, A., S., & Reiser, B. J. (2016). Epistemologies in Practice: Making scientific practices meaningful for students. *Journal of Research in Science Teaching*, *53*(7), 1082–1112.

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)*.

Kucirkova, N., Gerard, L., & Linn, M.C. (2021). Designing personalised instruction: A research and design framework. *British Journal of Educational Technology*, *52*(5), 1839-1861.

Gerard, L., Kidron, A. & Linn, M.C. (2019). Guiding collaborative revision of science explanations. *International Journal of Computer-Supported Collaborative Learning*, 14, 291–324.

Gerard, L. F., & Linn, M. C. (2016). Using automated scores of student essays to support teacher guidance in classroom inquiry. *Journal of Science Teacher Education*, *27*, 111–129.

Krajcik, J., Codere, S., Dahsah, C., Bayer, R., Mun, K. (2014). Planning instruction to meet the intent of the Next Generation Science Standards. *Journal of Science Teacher Education*, 25, 157–175.

Next Generation Science Standards Lead States (NGSS). (2013). *Next Generation Science Standards: For states, by states.* National Academies Press. Washington, DC.

Passonneau, R. J., Poddar, A., Gite, G., Krivokapic, A., Yang, Q., & Perin, D. (2018). Wise crowd content assessment and educational rubrics. *International Journal of Artificial Intelligence in Education*, *28*, 29-55.

Tansomboon, C., Gerard, L., Vitale, J., & Linn, M.C. (2017). Designing Automated Guidance to Promote Productive Revision of Science Explanations. *International Journal of Artificial Intelligence in Education*, 1-29.

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.

Acknowledgments

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