International Society of the Learning Sciences

AI-Human Partnership to Help Students Write Science Explanations

Sadhana Puntambekar, University of Wisconsin–Madison, puntambekar@education.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: In this paper, we present a case study of designing AI-human partnerships in a real-world context of science classrooms. We designed a classroom environment where AI technologies, teachers and peers worked synergistically to support students' writing in science. In addition to an NLP algorithm to automatically assess students' essays, we also designed (i) feedback that was easier for students to understand; (ii) participatory structures in the classroom focusing on reflection, peer review and discussion, and (iii) scaffolding by teachers to help students understand the feedback. Our results showed that students improved their written explanations, after receiving feedback and engaging in reflection activities. Our case study illustrates that **Augmented Intelligence** (USDoE, 2023), in which the strengths of AI complement the strengths of teachers and peers, while also overcoming the limitations of each, can provide multiple forms of support to foster learning and teaching.

Introduction

One of the key areas of AI in education is that of automated scoring of students' written work (Celik, et al., 2022). This is an important area where Natural Language Processing (NLP) techniques can help by providing immediate feedback to students (Gerard & Linn, 2022; Zhai et al., 2020), while at the same time supporting teachers who often have little time to provide detailed feedback on students' written work. In science writing, research suggests that students' often struggle with writing explanations (Berland & Reiser, 2009; McNeill & Krajcik, 2007). Scoring of essays is time consuming for teachers as they usually teach multiple classes with at least 25-30 students in each class. NLP technologies can help by providing timely feedback, but these technologies are only a first step in helping students with their written explanations. This is because merely making scores available to students is not adequate in classroom settings; developing feedback based on the automated scoring that is comprehensible to students is more important (Ke & Ng, 2019; Zhu et al., 2020). Even after students receive feedback, supporting the revision process for writing is not trivial (Tansomboon et al., 2017). Additionally, scoring is not always accurate, and students and teachers need to understand the fallacies of AI systems.

We have found these challenges in our work using NLP technologies in science classes. We used an NLP technology, PyrEval (Gao et al., 2018), to automatically assess students' essays in middle school science classrooms. PyrEval looks for main ideas in students' essays, determined by instructional goals in collaboration with teachers (Singh et al., 2022). It then gives students a list of the ideas they covered in their essays, and ones they missed. In our initial studies, we found that the AI-generated assessment was useful in helping students iteratively refine their ideas in a timely manner and improve their science explanations (Cang et al, 2023). However, we also found that students faced challenges revising their essays, due to potential inaccuracies or confusion about what to revise and how. Students often added superficial or disconnected ideas without integrating them into their original writing (Cang et al., 2023), or had trouble understanding the feedback itself, to be able to incorporate it into their revisions (Puntambekar et al., 2023).

Thus, there is the need for human-in-the-loop approaches to help address these challenges, so that students can learn how to revise, and improve their science writing. As proposed by Luckin and Cukurova (2019), learning sciences research has helped us understand how people learn, and provided insights about the processes of learning and teaching. Leveraging this knowledge as we design and use AI technologies in the classroom is key to helping students learn. Despite the rise the AI technologies in education (Crompton et al., 2022), there has been less work in the area of AI-human partnerships, a need that has been emphasized in a recent report from the US Department of Education (2023). It is especially important to understand how to design and implement such partnerships in classrooms, so that AI systems can positively impact teaching and learning.

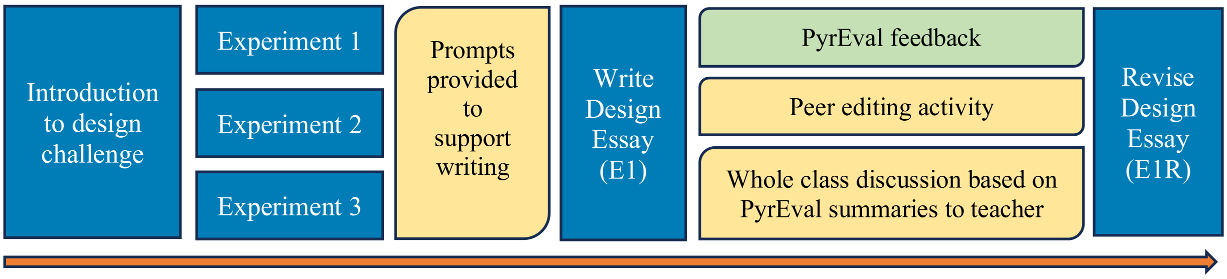
In this paper, we present a case study of designing AI-human partnership in a real-world context of science classrooms. Building on theories of distributed intelligence (Pea, 1997), distributed and synergistic scaffolding (Puntambekar, 2022; Tabak, 2004), and research in classroom participatory structures (e.g., Tabak & Baumgartner, 2004), we designed a classroom environment in which AI technologies, teachers and peers worked synergistically to support students' writing in science. To complement the automated assessment from PyrEval, we designed (i) feedback that was easier for students to understand; (ii) participatory structures in the classroom focusing on reflection, peer review, and discussion; and (iii) scaffolding by teachers to help students understand the feedback. We present our study as an example of AI-human partnership, including students' learning outcomes, and lessons we learned.

Methods

Study context and participants

Students (*N*=96) from five classes taught by one teacher in a school in the Midwestern US participated in this study. Students worked on a design-based physics unit on the science of roller coasters to learn about the relationships between motions, forces, and energy. They were challenged to design a roller coaster that was fun and safe based on physics, which they would need to explain in a design essay. To learn the science necessary to solve their challenge and gather evidence for their designs, students participated in multiple experiments using a virtual simulation, to manipulate variables (such as height of the track or mass of the car). They performed these investigations, recorded their findings, and wrote their design essays within a digital notebook. The digital notebook enabled students to engage in various activities such as, running a simulation, data generation and export, data review, data analysis and interpretation, and writing and submitting essays to get feedback from PyrEval. After students submitted their first essay, E1, it was sent to PyrEval for evaluation. Essays were evaluated by PyrEval for the presence of six, pre-determined science ideas and relationships, or content units, (CUs, described next) that should have been included in students' design essays. Students received the feedback in their digital notebooks, showing whether particular science ideas were present or not and an average accuracy level for each CU. Students were then given an opportunity to revise and resubmit their essays based on the automated feedback. This second essay was their revised essay, E1R. Figure 1 shows the sequence of activities in the unit.

Figure 1
Sequence of the roller coaster unit, showing the scaffolds designed to support students' writing


Introduction to design challenge
Experiment 1
Experiment 2
Experiment 3
Prompts provided to support writing
Write Design Essay (E1)
PyrEval feedback
Peer editing activity
Whole class discussion based on PyrEval summaries to teacher
Revise Design Essay (E1R)

Support from AI and classroom structures

In addition to providing students with feedback based on the automated assessment, we also introduced participatory structures (Tabak & Baumgartner, 2004) and routines, including peer review and reflection in groups, and teacher support to help students understand the feedback to revise their essays.

Automated essay analysis using PyrEval

We used PyrEval (Gao et al., 2018; Passonneau, et al., 2018) to evaluate student's essays. PyrEval works on a wise-crowd model, i.e., it assesses a student's essay for ideas by comparing it to a model constructed from essays from expert writers to the same prompts. PyrEval creates vector representations of ideas or content units (CUs) and assigns a higher weight to ideas more frequently present in the expert writing references. To evaluate a student's essay, PyrEval relies on meaning vectors extracted from the reference passages created by the wise crowd. By comparing the CUs in the student essay to the model CUs, PyrEval can assess the similarity of meaning between the student's writing and the model. We also included an average accuracy level for each of the CUs, i.e., how accurately PyrEval had (or had not) detected the idea in a development dataset. This information was provided to help students understand that the automatic feedback is not perfect. For example, if PyrEval determined a CU was present in an essay, but the accuracy based on a development dataset was low, students could be prompted to make sure it was truly there and revise accordingly.

Teachers as partners

The importance of participatory design with teachers as collaborators is seen in the development of curricula (Penuel, et al, 2007) in the design of AI technologies (Holstein & Aleven, 2022), as well as in our own work (Puntambekar et al., 2007). Throughout the process of development and implementation of PyrEval in classrooms, our participating teachers have collaborated as our partners. First, we conducted a week-long summer meeting with teachers in which they provided input during curriculum development, specifically on which CUs were important for students to learn, and for PyrEval to identify in students' writing. Second, after our first year of implementations, they provided feedback on how the automated assessment worked in their classrooms and what problems they encountered, and the issues that students had in interpreting the feedback. Last, and more importantly, in the summer meeting, we also helped teachers understand the factors that affect PyrEval's accuracy, and how they should discuss issues about AI with their students. For example, factors like sentence length, punctuation, and vague explanations of science ideas could affect PyrEval's accuracy (Puntambekar et al., 2023).

Support before writing an essay

To help students be mindful about their writing, we provided students with prompts before writing their essays. These prompts were based on the common problems that we found in students' writing such as lack of punctuation, repetition of ideas, being specific in explaining their ideas, and explaining the direction of relation between their ideas. Each student was given a laminated sheet with prompts that they could refer to anytime during their writing. We included prompts about writing, such as "don't forget punctuation!" and "add periods and commas where needed!" We also provided prompts about science writing. For example, one of the prompts was,

- It's important to specifically explain ideas. Always include the direction of the relation between ideas (increase/decrease, more/less, etc.).
    - Specific explanation: The higher the height the greater the potential energy.
    - Vague explanation: Height affects potential energy.

Reflection and peer review

Based on the findings from our previous study and teachers' feedback, we also introduced reflection and peer review for writing (Cho & MacArthur, 2010; Kollar & Fischer, 2010), so that students could understand the feedback, and reflect on *why* they may have received this feedback to help them make more targeted and substantive revisions. Students worked in groups for this activity. They were given three anonymous essays from the previous year, and PyrEval's assessment of each of the essays, which showed whether a particular main idea was identified by PyrEval or not. Prior to engaging in this activity in their small groups, in a whole class discussion, the teacher provided reflection prompts and guidance for the activity. For each of three essays students discussed and wrote: (i) whether an idea was correctly identified or missed by PyrEval; (ii) why they thought the idea was identified or missed; and (iii) for missed ideas, what suggestions they had to improve the writing. Students worked in their small groups to reflect on each of the examples, and discussed why some ideas were missed by PyrEval, even when they might have been mentioned in the essay. They discussed punctuation and grammar, and also whether the science ideas were written correctly. After students discussed ideas, they then wrote down suggestions for how to improve each of the essays. This activity was conducted over two class periods of 45 minutes each. At the end of the discussion in their groups, the teacher led a whole class discussion about students' suggestions and highlighted the strategies about writing in science.

Data

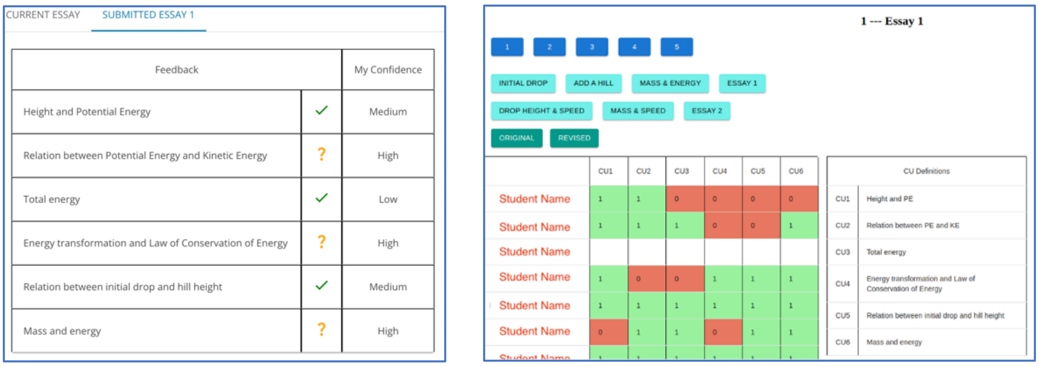
Automated essay scores generated by PyrEval

We used the data generated from PyrEval's assessment of students' essays, which showed whether students included important physics ideas, CUs, in their essays. Six CUs were identified as important to include in the design essays. PyrEval assessed whether these ideas were present (or not), generating a binary vector score for each CU for each student's essay. A score of 1 indicated that a particular CU was detected and a score of 0 meant the CU was not detected in the essay. An example of the feedback a student received can be seen in Figure 2 (left). This feedback was generated based on a vector score where only CUs 0, 2, and 4 were detected (e.g., [1,0,1,0,1,0]). As mentioned earlier, we also provided students with the accuracy level for each CU, which was based on how accurate the assessment by PyrEval was for a specific CU. We presented the PyrEval's accuracy level as "My confidence" level to help students understand the imperfect nature of AI and use PyrEval's feedback in critical and reflective ways. For our analyses, we used CU total scores for each essay by adding the total number of CUs mentioned on each essay. In the above example, a score of [1,0,1,0,0,1] would receive 3 as the total CU score for that essay. As mentioned previously, students wrote two essays, E1 and E1R. We examined the scores from these two essays, to understand changes from E1 to E1R. Additionally, teachers received a summary table of the scores of all their students by class, for each of the CUs (Figure 2, right).

Videos from classrooms

In addition to data from the evaluation by PyrEval, we also video recorded some student groups as they participated in the roller coaster unit in their classroom. We collected videos for every day of the instruction, from one target class. The target class was selected because it had the greatest number of students who consented to participate in the study, including the collection of audio and video. The target class was representative of the student population in that it had students with a range of academic abilities. For this paper, our focus was on the days in which students were asked to reflect on essays and engaged in the peer review activity. As mentioned earlier, this activity occurred over two class periods of 45 minutes each. Of the five groups in one class, we recorded two groups of students as they engaged in the reflection and peer review activity. These groups were selected based on input from the teacher as representative in terms of their academic performance as well as how well the teacher thought the students would discuss science ideas. These videos were transcribed resulting in 298 turns of talk in Group A and 251 turns of talk in Group B. Because we were only interested in the students' collaborations with one another, we removed all exchanges with the teacher, resulting in 291 turns of talk in Group A and 188 turns of talk in Group B. Two researchers independently viewed the videos and the transcripts to identify, discuss, and agree upon when students engaged in concrete discussion about the problems in the sample essays as well as concrete solutions. We then calculated percentages by dividing the turns of talk that contained this kind of talk by the total turns of talk (not including the teacher) and multiplying by 100. We also identified any other themes to characterize the nature of students' collaboration during the activity.

Figure 2
Scores from PyrEval for students, on left, and class summary received by teachers, on right



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

	CU1	CU2	CU3	CU4	CU5	CU6
Student Name	1	1	0	0	0	0
Student Name	1	1	1	0	0	1
Student Name						
Student Name	1	0	0	1	1	1
Student Name	1	1	1	1	1	1
Student Name	0	1	1	0	1	1
Student Name	1	1	1	1	1	1

CU Definitions	
CU1	Height and PE
CU2	Relation between PE and KE
CU3	Total energy
CU4	Energy transformation and Law of Conservation of Energy
CU5	Relation between initial drop and hill height
CU6	Mass and energy

Results

Automated essay analysis using PyrEval

To understand whether students' written explanations improved, we compared their CU total scores from E1 to E1R. Ninety-one of the 96 participating students wrote both E1 and E1R. Thus, this is the number of students we used for our paired samples *t*-test. We found that students significantly improved in the number of CUs they included in their essay from E1 (M(SD)=4.25(1.31)) to E1R (M(SD)=4.99(1.13)) ($t_{(90)}$ = -7.04, $p < .001$).

Reflection and peer review

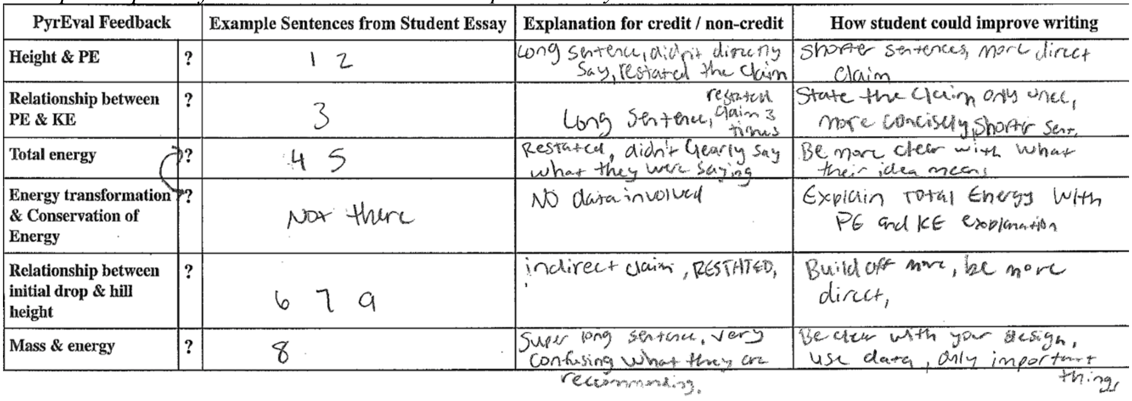
As mentioned earlier, we included a reflection and peer review activity so that students could reflect on essays and discuss in their groups what PyrEval identified correctly and what it did not, and think about ways in which the essay they were reviewing could be improved. This reflection activity was done in groups, after students wrote E1 and before they received feedback on their own essays. This was intended to help students understand the scores, and why sometimes it seemed that PyrEval might have given a specific score. Students discussed why PyrEval either identified or missed a CU. They identified problems in each essay and discussed and wrote their suggestions for improvement. Students deeply engaged with each other during this activity and observed several issues with the sample essays and made many suggestions for how the essays could be improved in their small groups. A sample of students' responses is shown in Figure 3. The figure illustrates how the student mentioned that the sentences in the sample essay were long, unclear, or indirectly stated the science relationships. They were also then able to make appropriate and substantive suggestions for how to fix the problem, as can be seen in the final column on the right.

We found that the two groups of students that we video recorded collaborated well during this activity and engaged in very minimal, if any off-task talk. Overall, they were focused on identifying several issues with the sample essays in relation to PyrEval's feedback and made many suggestions for how the essays could be improved in their small groups. Students discussed problems with the writing quality–e.g., whether the essays contained long sentences and restatements of the same idea. They also discussed solutions for these problems–e.g., suggestions to use shorter sentences, stating ideas more concisely, and avoiding repetition–which were all relevant and appropriate ideas for how to tackle revisions. In fact, discussions of concrete problems and solutions in the sample essays were present in about 25% of the turns of talk for Group A and 23% for Group B.

In looking deeper at the 25% of Group A's talk about concrete problems and suggestions for solutions, we found that about 15% of their talk focused on identifying problems with the writing; these were problems related to writing quality (~9%) and science content (~6%). Group A also had many turns of talk related to proposing solutions to the problems they identified (about 10%), and this talk was mostly evenly distributed between proposing solutions for fixing writing quality and science content (about 5% each). We also found that these conversations were present throughout their transcript.

Figure 3

Example response from students on how to improve essays



PyrEval Feedback		Example Sentences from Student Essay	Explanation for credit / non-credit	How student could improve writing
Height & PE	?	1 2	Long sentence, didn't directly Say, restated the claim	Shorter sentences, more direct claim
Relationship between PE & KE	?	3	Long Sentence, restated claim 3 times	State the claim only once, more concisely, Shorter Sen.
Total energy	?	4 5	Restated, didn't clearly say what they were saying	Be more clear with what their idea means
Energy transformation & Conservation of Energy	?	Not there	NO data involved	Explain Total Energy With PE and KE explanation
Relationship between initial drop & hill height	?	6 7 9	indirect claim, RESTATED,	Build off more, be more direct,
Mass & energy	?	8	Super long sentence, very confusing what they are recommending.	Be clear with your design, use data, only important things

While students in Group B had 23% of overall talk about problems and solutions, the proportion of problems and solutions discussed were different from Group A. The students in Group B focused more on the problems with the writing quality (10%) and science ideas (11%) in the sample essays (22% overall), rather than solutions for these problems (1% for writing quality solutions, and 0% for fixing science content). However, even though the students in Group B did not discuss many concrete solutions, they engaged in several in-depth conversations about science relationships, about 13% of their overall turns of talk. Further, these students also had several turns of talk where they reflected on problems in their own writing (based on issues with the sample essay).

Table 1 shows examples of students' discussions. In Excerpt 1 in Table 1, students were discussing the issues with a sentence about potential and kinetic energy (PE and KE) from one of the example essays. They first identified that the sentence was "very long" and that the same idea was stated multiple times. They then suggested that the writer should not restate the ideas so many times and be more concise in their writing. Similarly in Excerpt 2, students recognized that the sentences in the example essay were confusing. They then discussed concrete solutions to solve the problem, which further revealed that they understood what made the sentence confusing; the writing was unclear and it included extraneous information. Students also suggested including data, which is an important science practice. As these examples show, students reflected on the AI feedback, assessed whether the feedback made sense in the context of the writing, thought about the issues with the writing, and brainstormed specific ways that the writing could be improved. This provided them with important practice for approaching the feedback they received on their own essays, so they might be in a better position to reflect on and revise their own writing in meaningful ways. Additionally, and importantly, this activity helped students think about and discuss aspects of writing science explanations.

Table 1

Example student discussion about sentences from an essay

Excerpt 1	
Sentences from essay	The relationship between KE and PE is when there is no friction they transfer mechanical energy to each other. In other words, PE is mechanical energy and as

	it goes down the hill it gets smaller and smaller, and while it's doing that the KE's energy gets larger because Pe is transferring the mechanical energy to Ke.
Dialogue	S2: That's a very long sentence. He also just like… S3: That would be the relationship between KE and … S2: They also kind of like said the same thing twice. S1: Yeah, so it’s three, so that's three. Long sentence, restated… three times? S2: Yeah, I think so. The first sentence was once and then they did it three times, right? S1: So, state the claim only once, more concisely. S3: How they can improve? Hmm, stop restating… S1: Yeah, I said state it only once and like to be more concise. S3: And shorter sentences, too.
	Excerpt 2
Sentences from essay	For mass, I think 50kg because you could hold a lot of people but it doesn't matter because the mass doesn't affect the speed at all it affects energy because with more mass you get more PE which creates more energy. the hill should be 2.5 because the car will have enough energy to get up the hill and it will ensure it will be able to get through the ride.
Dialogue	S1: Yeah, very confusing what they are saying. S2: Don’t write lengthy sentences. S1: Okay, so be clear with your design, use data. S3: Don’t say stuff if it doesn’t matter.

Whole class discussion by teachers

In addition to students, we also provided teachers with summaries about the CUs students included in their essays for each of their classes (see Figure 2, right side). Teachers received a chart that showed the CUs identified by PyrEval so that they could quickly identify which CUs were covered by most of their students and which were not. This gave the teacher an opportunity to get information on which key ideas students missed, so she could use this information in a whole class discussion to further scaffold students’ revisions. For this discussion, the teacher displayed information about the ideas most students from the class included in their essays and which they did not, as assessed by PyrEval, on the display at the front of the classroom. Then, as illustrated in Table 2, she discussed that, first, most students seemed to have understood the relationship between height and potential energy. However, she also mentioned that several students missed the connections between total energy and PE and KE, so those ideas needed more work. Second, she also noticed that students had discussed energy transformation, but they did not always make the connection to conservation of energy.

The two issues the teacher mentioned were related, and signaled a lack of understanding on a general principle, that of conservation of energy, that students needed to learn. By seeing the data from the whole class generated by PyrEval, the teacher could quickly see where support might be needed for students to understand these science ideas. This discussion about the science ideas students had included or missed was another layer of scaffolding, since it was based on the data provided by PyrEval about the trends in that particular class.

Table 2
Excerpt from the teacher discussing patterns in CU coverage

So, using that summary grid, I noticed…vast majority talked about initial drop height versus hill height and the relationship between height and PE. So, looking across the board, almost everyone got the green checkmark for these two topics. However, looking at the other end of the scale. Total energy and relationship between PE and KE were the ones that had the most question marks… you may need to do some revising to make it more clear, right? I noticed that that whole idea of energy transformation was in there, perhaps needing to do a better job of connecting energy transformation to conservation of energy to make that into our best topics category.

Discussion

The case study we discussed in this paper is an example of a successful AI-human partnership in the real world context of middle school science learning. Our case study illustrates that Augmented Intelligence (U.S. DoE, 2023), in which the strengths of AI complement the strengths of teachers and peers, while also overcoming the limitations of each, can successfully support learning and teaching. In our study, PyrEval gave students immediate assessment on their essays, which is hard for teachers to do in a timely manner. But while the assessment might be automated, generating feedback that is comprehensible and actionable based on the output scores is challenging. Further, even after students receive feedback, helping them understand *what* to revise and *how* to revise is important. Our results showed that students improved in the number of key ideas that they mentioned in their revised essays, which was likely fostered by students receiving the automatic feedback in combination with participating multiple activities that helped them to reflect on how to use the feedback to revise. For example, students discussed how to write a good science explanation, by reflecting on example essays and thinking about strategies to improve them. Further, our results showed that the teacher was able to use summary scores from PyrEval to understand gaps in students' learning and address them during instruction to further help students to reflect on what they would likely need to revise in their essays. Additionally, other scaffolds, such as the prompts we provided to students before they wrote their first essay helped them understand how to write, and also served as guidance for reflection on the example essays. Thus, the automated scores from PyrEval served as a springboard to start a conversation about writing explanations in science. The prompts, classroom structures such as peer review and reflection, and teacher facilitation worked synergistically to support students by helping them understand key aspects of science writing, such as being specific, including data, and including directionality of relations between science concepts.

Lessons learned from learning sciences research, enabled us to design the classroom learning environment in which AI played a key role, but not the only role. Our study capitalized on decades of research on how people learn–specifically on providing prompts to support science learning (Martin et al., 2019; Reiser, 2004), collaborative learning, peer review and reflection for written explanations (Berland et al., 2016; Gerard et al., 2019), and on teacher scaffolding (e.g., van de Pol et al., 2010). Rather than starting with a technology tool, our research started with a problem that we wanted to address, and with knowledge based on learning sciences research on ways to support learning, and used AI to fill a critical gap, that of real-time feedback on writing.

Our results contribute to the body of research in the learning sciences examining how teachers, peers and AI can play complementary roles to support learning (Gerard et al., 2019; Holstein & Aleven, 2022). We have developed ways in which the automated feedback could be integrated in classroom structures so that teachers and peers are part of the feedback and revision process, rather than relying on the feedback from AI alone. We found that PyrEval can provide timely and immediate feedback to students and provide feedback multiple times without burdening the teacher. Students working collaboratively can support each other by reflecting on and discussing the feedback. And teachers can model how to revise and design instruction based on the trends provided to them through PyrEval. In future work, our plan is to understand how multiple teachers use the feedback and revision processes, and examine learning in classrooms in a variety of settings.

References

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

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.

Cang, X., Gnesdilow, D., & Puntambekar, S. (2023). Revisions in scientific explanations using automated feedback. *Proceedings of the 17th International Conference of the Learning Sciences (ICLS)*. International Society of the Learning Sciences.

Celik, I., Dindar, M., Muukkonen, H., & Järvelä, S. (2022). The promises and challenges of artificial intelligence for teachers: A systematic review of research. *TechTrends*, *66*(4), 616-630.

Cho, K., & MacArthur, C. (2010). Student revision with peer and expert reviewing. *Learning and Instruction*, *20*(4), 328-338.

Crompton, H., Jones, M. V., & Burke, D. (2022). Affordances and challenges of artificial intelligence in K-12 education: A systematic review. *Journal of Research on Technology in Education*, 1-21.

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)*.

Gerard, L. & Linn, M.C. (2022). Computer-based guidance to support student revision of their science explanations. *Computers & Education*, 176, 304-351.

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

Holstein, K., & Aleven, V. (2022). Designing for human–AI complementarity in K-12 education. *AI Magazine*, *43*(2), 239-248.

Ke, Z., & Ng, V. (2019, August). Automated essay scoring: A survey of the state of the art. In *IJCAI* (Vol. 19, pp. 6300-6308).

Kollar, I., & Fischer, F. (2010). Peer assessment as collaborative learning: A cognitive perspective. *Learning and Instruction*, *20*(4), 344-348.

Luckin, R., & Cukurova, M. (2019). Designing educational technologies in the age of AI: A learning sciences-driven approach. *British Journal of Educational Technology*, *50*(6), 2824-2838.

Martin, N. D., Dornfeld Tissenbaum, C., Gnesdilow, D., & Puntambekar, S. (2019). Fading distributed scaffolds: The importance of complementarity between teacher and material scaffolds. *Instructional Science*, *47*, 69-98.

McNeill, K. L., & Krajcik, J. (2007). Middle school students' use of appropriate and inappropriate evidence in writing scientific explanations. In M. C. Lovett & P. Shah (Eds.), *Thinking with data* (pp. 233–265). Lawrence Erlbaum Associates Publishers.

Pea, R. D. (1997). Practices of distributed intelligence and designs for education. In G. Salomon (Ed.), *Distributed cognitions: Psychological and educational considerations* (pp. 47-87).

Penuel, W. R., Roschelle, J., & Shechtman, N. (2007). Designing formative assessment software with teachers: An analysis of the co-design process. *Research and Practice in Technology Enhanced Learning*, *2*(01), 51-74.

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. *Proceedings of the 17th International Conference of the Learning Sciences (ICLS)*. International Society of the Learning Sciences.

Puntambekar, S., Stylianou, A., & Goldstein, J. (2007). Comparing classroom enactments of an inquiry curriculum: Lessons learned from two teachers. *Journal of the Learning Sciences*, *16*(1), 81-130.

Reiser, B. J. (2004). Scaffolding complex learning: The mechanisms of structuring and problematizing student work. *Journal of the Learning Sciences, 13*(3), 273–304.

Singh, P., Gnesdilow, D., Cang, X., Baker, S., Goss, W., Kim, C., Passonneau, R., & Puntambekar, S. (2022). Design of Real-time Scaffolding of Middle School Science Writing Using Automated Techniques. In C. Chinn, E. Tan, C. Chan & Y. Kali. (Eds.), *International Collaboration toward Educational Innovation for All: Overarching Research, Development, and Practices: ICLS 2022*, (pp. 1521-1524).

Tabak, I. (2004). Synergy: A complement to emerging patterns of distributed scaffolding. *Journal of the Learning Sciences, 13*(3), 305–335

Tabak, I., & Baumgartner, E. (2004). The teacher as partner: Exploring participant structures, symmetry, and identity work in scaffolding. In *Investigating participant structures in the context of science instruction* (pp. 393-429). Routledge.

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.

U.S. Department of Education, Office of Educational Technology. (2023). *Artificial intelligence and future of teaching and learning: Insights and recommendations*. Washington DC.

van de Pol, J., Volman, M., & Beishuizen, J. (2010). Scaffolding in teacher–student interaction: A decade of research. *Educational Psychology Review*, *22*(3), 271-296.

Zhai, X., C Haudek, K., Shi, L., H Nehm, R., & Urban-Lurain, M. (2020). From substitution to redefinition: A framework of machine learning-based science assessment. *Journal of Research in Science Teaching*, *57*(9), 1430-1459. https://doi.org/10.1002/tea.21658.

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. https://doi. org/10.1016/j.compedu.2019.103668

Acknowledgments

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