International Society of the Learning Sciences

# Revisions in Scientific Explanations Using Automated Feedback

Xuesong Cang, Dana Gnesdilow, Sadhana Puntambekar
xcang@wisc.edu, gnesdilow@wisc.edu, puntambekar@wisc.edu
University of Wisconsin, Madison

**Abstract:** Writing and revising scientific explanations helps students integrate disparate scientific ideas into a cohesive understanding of science. Natural language processing technologies can help assess students' writing and give corresponding feedback, which supports their writing and revision of their scientific ideas. However, the feedback is not always helpful to students. Our study investigated 241 middle school students' a) use of feedback, b) how it affected their revisions, and c) how these factors affected students' writing improvement. We found that students made more content-related revisions when they used feedback. Making content-related revisions also assisted students in improving their writing. But students still found it difficult to make integrated revisions and did not use feedback often. Additional support to assist students to understand and use feedback, especially for students with limited science knowledge, is needed.

## Introduction

Writing science explanations is an integral part of learning and doing science (NGSS, Lead States, 2013). Students need to explore, understand, and explain why scientific phenomena happen using scientific ideas. Writing scientific explanations provides opportunities for students to integrate their disparate scientific ideas into more cohesive and deeper understandings of science topics (Braaten & Windschitl, 2011; Linn, 2006). Though it is often challenging for students, making revisions of science writing can prompt them to connect their initial ideas to new ideas, see the connections between scientific ideas, and strengthen their understanding of science (Linn, 2006; Tansomboon et al., 2017). The development of skills in science writing and revisions of writing has been found to benefit students' long-term science learning (Rivard, 1994). Despite the importance of writing in science and making revisions, students usually get minimal support in writing and revising their science ideas in the classroom. This is likely because it is challenging for teachers to read and provide individualized, constructive feedback in a timely manner to students given limited class time and the number of students teachers have (Gerard et al., 2022). In recent times, Natural Language Processing (NLP) technologies are being used to provide timely and detailed feedback.

NLP tools can provide feedback to support students in understanding the gap between what they have written as well as what they have missed in their explanations (Hattie & Timperley, 2007; Roscoe et al., 2015). In this way, the automated feedback, generated from natural language processing technologies, can play the role of more "knowledgeable others" from a sociocultural perspective to support students in making revisions (Vygotsky, 1978). However, students often face difficulties in using automated feedback. Some students find it challenging to understand computer guidance and trust the feedback. As a consequence, students ignore it (Zhu et al., 2017). Other students may struggle to make revisions based on automatic feedback because they do not understand the science well (Zhu et al., 2020). Further, students' revisions of science writing are influenced by writing practices in school, where science revisions are most often viewed as accumulating ideas (Gerard et al., 2022). Students and teachers may have different understandings about what it means to engage in revisions of science writing and may take different approaches. As a result, automated feedback is not always helpful in supporting students to make in-depth revisions (Shute, 2008). This may, at least partially, explain why students tend to either simply add new, relevant ideas without integrating them into previous writing or elaborate on the existing ideas repeatedly without modifying their initial writing (Gerard et al., 2016). Further, some other students choose not to revise their ideas when they are supported with automatic feedback.

Prior studies have provided little information about how students use feedback based on automated assessments and if the use of feedback influences the types of revisions that students engage in. It is important for researchers to understand how students' revise their scientific explanations so they can enhance the effectiveness of automated assessment and feedback as well as better design supports that students need to engage meaningfully in writing and revising their science ideas (Tansomboon et al, 2017). Therefore, we need to understand how students actually engage in revisions after getting feedback, and whether students' revisions improve their writing (Lee et al., 2019, 2021). The goal of this study was to understand the ways in which students revised their scientific explanations in an essay based on automated feedback provided by a natural language processing software, called PyrEval (Gau et al. 2018; Singh et al., 2022). Our research questions were as follows:1)How do students revise

their explanations based on automated assessment and feedback?; 2)To what extent do students use feedback in their revisions? and 3)What are the differences in the types of revisions students make?

## Methods

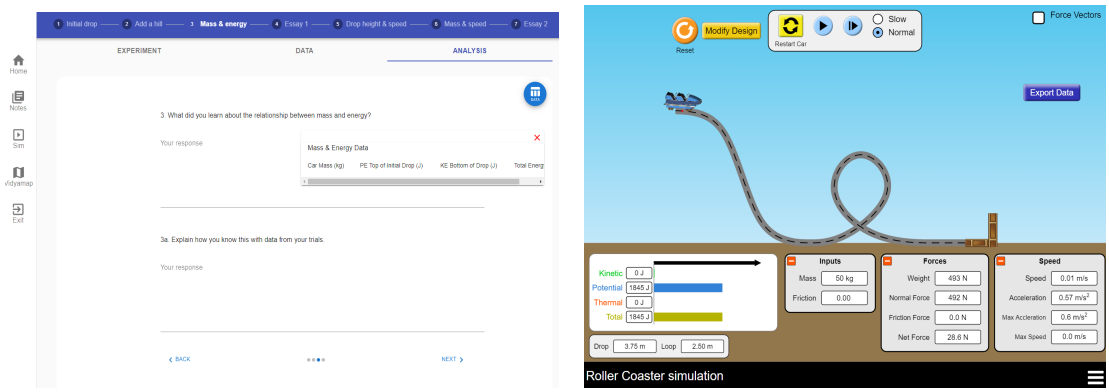
### Participants and Study Context

Three 8th grade science teachers and their 241 students ($N_{T1}$=90; $N_{T2}$=67; $N_{T3}$=84) from two semi-rural public-school districts in the midwestern United States participated in this study. The students for whom we did not have full data were excluded. Students from Teacher one's (T1) classes were in a different school district from Teachers two and three's (T2 and T3). All teachers received the same professional development before the implementation of a physics unit in their classes. The teachers' professional development was related to the physics unit and use of our NLP automatic assessment to provide students with feedback for their writing.

During the design-based physics unit, students designed a safe and fun roller coaster based on what they learned about physics during the unit. The unit was taught over approximately fifteen 45-minute class periods. Throughout the unit, students used a digital notebook and conducted virtual experiments using a roller coaster simulation (Figure 1), and recorded data based on their experiments in the simulation. Students learned crosscutting concepts about energy and energy transfer within a roller coaster system (i.e., The Law of Conservation of Energy, potential energy and kinetic energy). Students wrote essays to explain their roller coaster design based on the science they learned during the unit and received feedback on their essays from our NLP system, PyrEval (described below). We provided students with prompts for their writing to help them understand which science ideas and relationships they should include in their essays, such as explanations about how height influences potential energy, or how energy transfers as the roller coaster car moves down the initial drop. The sequence of the unit was as follows: students a) were introduced to the roller coaster design challenge; b) conducted five virtual experiments to learn relationships between important science concepts that would help them to design a fun and safe roller coaster; c) wrote their roller coaster essays; e) received feedback from PyrEval the day after writing their essays; and e) revised essays.

**Figure 1**

*Science notebook (left) and simulation (right)*




The NLP software we used to assess and provide feedback on students' essays, PyrEval, was developed to identify weighted vectors of key content ideas (CUs) and relationships that students should include in their essays using a wise-crowd method (Gau et al. 2018; Singh et al., 2022). PyrEval parsed students' writing into propositions and assessed whether each proposition was a fit with each of 15 key content units that we identified as important for students to include in their essays (See Table 1 for some of the important CUs). A binary score of 1 or 0 was provided in PyrEval logs, presence of an idea was marked as a 1 and an absence was marked as 0. Students got feedback based on whether PyrEval identified important CUs in their essay. If PyrEval did not find any one of these 4 most highly weighted CUs, that were grouped into themes from the original 15 CUs (See Table 1), it provided students with feedback for improvement. Students were also provided with positive feedback if any of these CUs were detected. The feedback consisted of high-level, general statements and questions aimed at getting students to reflect on which concepts they explained and where they could improve based on PyrEval's assessment of their writing. The feedback that students got was similar to this example:

"You did a great job explaining Law of conservation of energy! You also wrote that the initial drop height should be higher than the hill height. Now, can you explain how PE at the top and KE at the bottom are related? Also write about how mass affects PE and KE."

**Table 1**
*Most Highly Weighted 4 CUs PyrEval Used to Generate Feedback*

| CU# | Science Idea / Relationship |
|---|---|
| CU0 | Potential and kinetic energy transform back and forth as the car moves and changes height |
| CU1 | Greater mass mean greater energy |
| CU2 | Explaining the Law of Conservation of Energy |
| CU3 | Initial drop must be higher than the hill to have enough energy to make it to the end of the ride |

## Data Sources and Measurement

Number of CUs per Essay: Students' essays were analyzed automatically using PyrEval. As mentioned above, up to 15 CUs could be identified by PyrEval. We analyzed the total number of CUs that PyrEval identified in students' essays as the *initial CU score* (from the final essay) and *revised CU score* (from revised final essays) to understand if PyrEval detected more key CUs in students' revised essays. At the same time, we also generated a *CU change score,* which was calculated by subtracting the *initial CU score* by the *revised CU score*. The *CU change score* shows the improvement of content units. The change score could be either positive or negative, depending on whether more or fewer CUs were identified in the revised final essays.

Revised or Not: Some students revised their essays while others did not revise their essays. We developed a binary code, *Revised or Not*, to capture if students revised their final essays.

Types of Revisions: We examined the data from both inductive and deductive perspectives. We first used the categories of revisions identified in Gerard et al. (2016) as a lens for our data and then inductively developed our coding scheme to fully capture the types of revisions that were in our dataset. Thus, we developed a binary coding scheme that captured four types of revisions that students engaged in: (1) *surface-level revisions,* making changes in spelling or word choice; (2) *added similar content,* repeating an existing science ideas or relationship that was already in their essay; (3) *added new content*, including new science ideas or relationships that were not in their initial essay; and (4) *integrated revisions,* reformulating ideas to improve the science ideas and relationships that were already written in their essay (See Table 2). Students could engage in multiple forms of revisions and receive multiple codes.

**Table 2**
*Examples of Types of Revisions*

| Revisions | Examples | Notes |
|---|---|---|
| Surface level revisions | **Feedback:** "You did a great job relating height with PE and KE! … can you explain how mass affects PE and KE?"<br>**Final Essay**: "when the potential energy went up the kintetic would to affecting in the speed being higher"<br>**Revised Final Essay**: "when the potential energy went up the **kinetic** would to affecting in the speed being higher" | Student corrected spelling of "kinetic" in the revised essays but did not address the feedback that suggested explaining how mass affects PE and KE. |
| Added similar content | **Feedback:** "…Can you explain how height affects PE and KE while explaining Law of conservation of energy? …"<br>**Final Essay:** "I believe that the initial drop should be 3 because it's fun and safe, it also will give a higher amount of PE which helps it have enough energy to go up the hill."<br>**Revised Final Essay:** "I believe that the initial drop should be 3 **meters** because it's fun and safe, it also will give a higher amount of PE which helps it have enough energy to go up the hill. **When you have a higher hill height, you get more PE because there is more potential for energy because the hill is higher.**" | Student explained height affects PE in the final essay without stating a direct relationship. Then the student added a sentence which explained a higher height means more PE in their revision, which made the writing more precise. |

| | | |
|---|---|---|
| Added new content | **Feedback:** "You did a great job explaining Law of conservation of energy! …Also write about how mass affects PE and KE."<br>**Final Essay:** NO writing about how mass affects PE and KE<br>**Revised Final Essay:** "**On the other hand though, mass effects potential and kinetic energy. The heavier the mass is, the more potential and kinetic energy is created**." | Student did not explain how mass affects PE and KE in the initial essay. Then added a sentence about it in their revision, based on the feedback. |
| Integrated revisions | **Feedback**: "… Can you explain how height affects PE and KE while explaining Law of conservation of energy? …"<br>**Final Essay:** "When we have 4ft at the starting drop, it makes the KE at the bottom the same as the PE at the top because all the PE is transferred into the KE at the bottom because the law of conservation of energy states that energy can be transferred but not created nor destroyed"<br>**Revised Final Essay:** "When we have a **4ft higher height** at the starting drop height, it makes the **PE at the top greater as well as the KE at the bottom** because all the PE is transferred into the KE at the bottom because the law of conservation of energy states that energy can be transferred but not created nor destroyed" | Based on the feedback, the students explained that height is directly related to potential energy by explaining "4 ft higher height…make the PE at the top greater as well the KE at the bottom" in the revised essays, thereby reformulating their previous writing. |

*Students' revisions have been bolded for emphasis.

Use of Automated Feedback: Though students received automated feedback to help them revise their writing, students did not always use or follow it. We developed a second coding scheme to capture whether students used the feedback they received from PyrEval. Students either (1) used the feedback by writing about the science concepts that they were asked to address, or (2) did not use the feedback. This coding was binary as well, with students receiving a 1 for using feedback and 0 for not using it. Each student could only have one type of code. Interrater agreement was established for both sets of coding categories. For each set of codes two researchers independently coded 15% of all revised essays and achieved almost perfect agreement (Stemler, 2001) on the types of revisions and the use of automated feedback codes (Kappas of .826 and .817, respectively). All discrepancies were resolved through discussion and the two researchers coded the remainder of the data.

## Data Analyses

We identified how many students revised their final essay. Of the 241 students who wrote the final essay, 87 of them revised their essays in some way. From here, we generated two sets of data: (1) a full dataset containing 241 students, and (2) a subset of the full data containing the 87 students who revised the essays. These datasets were used in different analyses described next.

First, we wanted to compare the *initial CU scores*, as well as the *revised CU scores* between students who revied and did not revise. To do this, we conducted an independent two-sample t-test using the full dataset. We also wanted to understand whether making revisions resulted in different levels of improvement in students' CU scores. For this, we ran a one-way between subjects Analysis of Variance (ANOVA) to compare the *CU change scores* between essays between students who revised or did not revise their essays. Second, for the eighty-seven students who revised their final essay, we wanted to understand how they used the feedback as well as the types of revisions they made. We calculated the percentage of students who engaged in each category of the types of revisions and the use of feedback. Beyond providing descriptions of students' use of the feedback and types of revisions they made, we also wanted to understand how the types of revisions may have been related to students' use of automated feedback. We conducted four pairs of chi-squared tests of homogeneity for each type of revisions and use of feedback: (1) surface-level revisions previous writings and the use of feedback, (2) added similar content and the use of feedback, (3) added new content and the use of feedback, and (4) integrated revisions and the use of feedback. Third, we wanted to more deeply explore if other factors along with the types of revisions and the use of feedback may have influenced the number of CUs that students mentioned in their revised final essays. To do this, we conducted a stepwise regression analysis, which included three fitted models.

## Results

### Comparison in the Number of CUs in essays

We first conducted an independent t-test using students' *initial CU score* and *Revised CU scores* to compare any differences between students who did or did not revise their essay. This test was appropriate since the CU score datasets were both normally distributed. Based on PyrEval's assessment of students' essays, we found that students who revised their essay included significantly more CUs in their final essays, $t_{(239)} = 2.00$, $p = .05$.

Similar results were found for the revised final essays; the 87 students who revised had significantly higher *revised CU scores* than students who did not revise, $t_{(239)} = 2.76$, $p = .005$ (see Table 3). Further, we conducted an ANOVA to examine the *CU change score* between students who revised or not. The results showed that students who made revisions had significantly higher *CU change scores* than students who did not revise their essays ($F_{(1, 239)} = 7.95$, $p < .001$).

**Table 3**
*Mean of CU scores, t test and ANOVA tests*

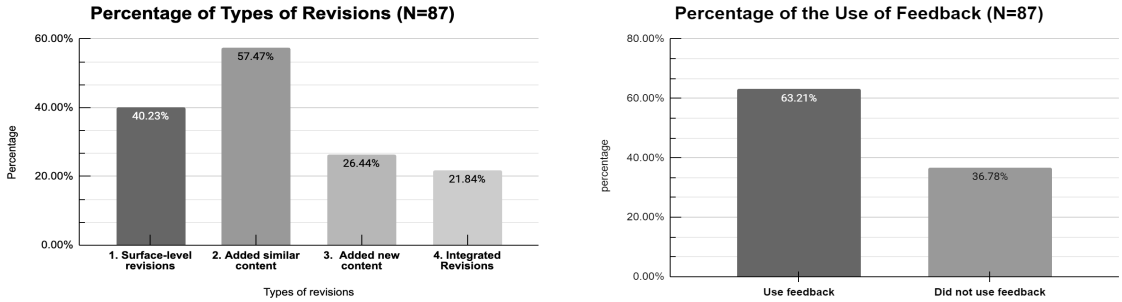
| | N | *Initial CU score* | *Revised CU score* | *CU change scores* |
|---|---|---|---|---|
| Revised | 87 | 4.93 (2.64) | 5.26 (2.76) | .33 (0.81) |
| Did not revise | 154 | 4.16 (3.00) | 4.16 (3.00) | 0 (0) |
| t-test or ANOVA | 241 | $t_{(239)} = 2.00$<br>$p = .05$ [1] | $t_{(239)} = 2.82$<br>$p = .005$ [2] | $F_{(1, 239)} = 25.75$<br>$p < .001$ [3] |

Note: [1] and [2] are t-test; [3] is ANOVA test

## Understanding Students' Revision Behaviors

To understand how students revised their essay, we examined two dimensions: (1) the *types of revisions* that students engaged in (e.g., surface-level revisions, adding similar or different ideas etc.), and (2) *use of feedback* (e.g., used the feedback or not). For the types of revisions, we found that students most often revised by adding similar content (57.47%) or by making surface-level revisions (40.23%). However, fewer students added new content (26.44%) or made integrated revisions in their essays (21.84%) (See Figure 2a). For the use of feedback, we found that more students used the automated feedback (63.21%) than students who did not (36.78%) in making their revisions (Figure 2b).

**Figures 2 a & b.**
*Percentage of types of revisions (left) and use of feedback (right)*




Relationship between types of revisions and use of feedback: To examine if there was a relationship between whether students used the feedback from PyrEval and the types of revisions they made, we ran four pairs of chi-square tests (Table 4). We found that there were statistically significant differences for two of the four tests: surface-level revisions and the use of feedback ($X^2_{(1, n=87)} = 11.96$, $p < .001$), and adding similar content and the use of feedback ($X^2_{(1, n=87)} = 15.99$, $p < .001$). There were no significant differences for adding new content or making integrated revisions and the use of feedback. We found students who did *not* use feedback were significantly more likely to make surface-level revisions than students who used or followed the feedback. We also found that students who added similar ideas were significantly more likely to have used the feedback than students who did not use feedback to inform their revisions.

**Table 4**
*Chi-square Omnibus Tests*

| | Surface-level | Added similar | Added new | Integrated |
|---|---|---|---|---|

| | revisions | content | content | revisions |
|---|---|---|---|---|
| The use of feedback | $X^2_{(1, n=87)} = 11.96$ $p < .001$ ** | $X^2_{(1, n=87)} = 15.99$ $p < .001$ ** | $X^2_{(1, n=87)} = .98$ $p = .32$ | $X^2_{(1, n=87)} = 1.79$ $p = .18$ |

Notes: ** is significant at .05 level and *** is significant at .01 level

### Exploring Factors that Influence Students' Science Writing

As prior studies have shown, engaging students in revising their science writing is an important practice to help them improve their science writing and learning. To understand the factors that might have affected students' scientific writings (*CU change score* as the dependent variable), we conducted a stepwise regression analysis by using factors including (i) types of revisions (*surface-level revisions, added similar content, added new content, integrated revisions*), (ii) use of feedback, and (iii) teachers (see Table 5).

We first conducted a multiple linear regression to better understand to what extent the variation in students' increased CU scores could be explained by four types of revisions. A significant regression equation was found ($F_{(4, 82)} = 2.38$, $p = .05$), with an $R^2$ of .10. Students' predicted *CU change score* was equal to - .11 + .28*(*surface-level revisions revisions*) +.54*(*Added similar content*) + .30*(*added new content*) -.29*(*integrated revisions*), where all the independent variables were coded as: 1 = presence of the type of revisions, 0 = absent. However, only one category of types of revisions, *added similar content,* was a statistically significant predictor. Students' *CU change score* increased .54 if students *added similar content* in the revisions.

Next, we examined whether the *use of feedback* was a predictor that explains CU scores in students' revised essays. For model 2, we excluded the non-significant predictors (*surface-level revisions, added new content, integrated revisions*) and added the *use of feedback* to predict the *CU change score*. But we did not find that model 2 was significant. Neither *added similar content* (one type of revisions) nor the *use of feedback* were significant predictors of *CU change score.* Based on the results from model 1 and model 2, we could see that only one type of revision, *added similar content,* was a predictor that explained students' CU change scores. In model 3, we further explore one more factor, teachers, and excluded non-significant predictor, use of feedback. Students' predicted *CU change score* were significantly predicted by this model 3 with an $R^2$ of .16. However, the factor of teachers is the only significant predictor.

**Table 5**

*Regression models to predict improvement in scientific explanations*

| | Outcome | Predictors | Regression Model results |
|---|---|---|---|
| Model 1 | CU change score | surface-level revisions;<br>Added similar content; **<br>Added new content;<br>Integrated revisions | $F_{(4, 82)} = 2.38$; $p = .05$; **<br>$R^2 = .10$; Adjusted $R^2 = .06$ |
| Model 2 | CU change score | Added similar content;<br>Use of feedback | $F_{(2, 84)} = 2.74$; $p = .07$.<br>$R^2 = .06$; Adjusted $R^2 = .04$ |
| Model 3 | CU change score | Added similar content;<br>Teachers *** | $F_{(3, 83)} = 5.24$; $p = .002$.<br>$R^2 = .16$; Adjusted $R^2 = .13$ |

Notes: ** is significant at .05 level and *** is significant at .01 level

## Discussion

While researchers have emphasized that writing scientific explanations and making revisions can help students integrate and connect scientific ideas (Braaten & Windschitl, 2011; Linn, 2006), these competencies are challenging to middle school students (Tansomboon et al., 2017). Students possess limited knowledge about how to revise their ideas to improve their writing (Hattie & Timperley, 2007). Further, teachers often do not have the time to provide students with detailed feedback to help them improve their writing (Gerard & Linn, 2022). To better support students and teachers, many researchers have developed technologies to automatically assess and provide feedback to students to help them improve their writing (Gernard et al., 2016). In this study, we wanted to know more about (1) the types of revisions students made to their scientific explanations, (2) the use of

feedback, (3) the relationship between these two aspects, and (4) how revisions may have led to improvements in writing to inform our work in better designing scaffolds to support their writing.

Our exploration of the revisions based on feedback and if the writing improvement showed that: (1) students were more likely to simply *add similar content* to their original writing and make surface-level revisions as opposed to the integrated revision. These results are aligned with prior studies that students find it challenging to: (1) see the gap between what they have written and what is missing, (2) connect scientific ideas, which means they tend to revise as if the science ideas are isolated or disconnected from what they wrote originally (Hattie & Timperley, 2007; Gerard & Linn, 2022). But we found that *added similar content* is an important type of revision that resulted in the detection of more content units in the exploration of how making revisions leads to improvement of writing by conducting multiple regression analyses. Even though students did not integrate these scientific ideas into their original writing, when they added similar content, they improved their writing by explaining ideas more specifically so that PyrEval could better detect the content units. Despite prior studies suggesting that integrated revisions aid in science learning (Gerard et al., 2016; Gerard & Linn, 2002), our findings demonstrated that reflecting on and revising scientific ideas through adding similar content can also lead to improvement in students' writing. Revisiting and revising explanations based on feedback likely helped students to better integrate their ideas, making them more cohesive (Braaten & Windschitl, 2011; Linn, 2006). Additional support is required to assist students in comprehending how to integrate their original ideas, rather than simply concatenating similar ideas. Further, it could also be the case that since so few students made integrated revisions, there was not enough power to detect its effect on improvements in students' essays related to the number of content units detected by PyrEval after they made revisions.

We investigated feedback use and found that many students did not use it, aligning with prior studies showing challenges in understanding and addressing it (Zhu et al., 2020). Some students may have simply not followed the feedback because they didn't know what to do, there may have been too much to address, or they simply did not know how to improve their ideas. We further explored how the use of feedback may have been related to the types of revisions students made, which has not been widely studied thus far. Even though prior studies indicated that some students tended to have superficial revisions with the support of automated feedback (Gernard et al., 2016; Shute, 2008), our study showed that the use of feedback did help students to make more content-related (i.e., added similar or new content), instead of surface-level revisions (i.e., fixed spelling). But using the feedback to make revisions was less effective in making integrated revisions, as few students did so in our study. Additional support can be designed to help students to reflect on what they have written in reference to the automatic feedback and assess whether they have explained their science ideas and, if not, fix what they have written in an integrated way.

Our results showed that students improved their writing based on the assessment of CUs from PyrEval. It indicates that making revisions helped students to improve their writing. However, the fact that students who did not revise had lower initial CUs scores in their original final essays may indicate that they may have not addressed the automated feedback because they had limited prior knowledge, which was also a finding from Zhu et al. (2017). This may indicate that students struggled to clearly write about the science ideas in a way that could be detected by PyrEval, because perhaps they did not understand the ideas very well. In addition, we also found that the teacher was a significant factor that predicted students' improved writing. It indicates that helping teachers to better support students to understand the feedback and make revisions is essential (Shute, 2008). This can be done at a whole class, group, and individual level. For example, before students revise, teachers can discuss this process more deeply in a whole class discussion. This kind of scaffolding may be essential especially for students with lower prior knowledge (Gerard & Linn, 2022). Though it is challenging to write and revise scientific explanations, the use of automated feedback with teachers' facilitation provided students in our study with the opportunity to practice writing scientific explanations and making revisions in the classroom, which is rare (Gerard et al., 2019). Our findings show a potential for helping a larger number of students to engage in this complex practice in science learning by using automated feedback.

## Conclusions

Writing and revising scientific explanations helps students to strengthen their understanding of science. However, there are many factors that impede the implementation of this practice, such as the number of students that one teacher needs to give feedback in the classroom or the limited knowledge students may have about science and / or about how to revise their science writing. We provided students with the automated feedback by using a NLP software, PyrEval, to provide feedback to help students to revise their writing. Our investigation of the types of revisions demonstrated that the automated feedback could positively shape students' approaches to revising their writing (i.e., focus more often on the scientific content and have less superficial level of revisions). However, the effectiveness of the tool still needs to be improved because it was less successful in getting students to make

integrated revisions, which requires the students to reformulate their original writing. Based on our findings that improvements in students' scientific writing were associated with certain types of revisions (added similar content), we could focus more on helping students understand how to use the feedback to meaningfully revise and integrate their science ideas to further improve their scientific writing and learning.

## References


Braaten, M., & Windschitl, M. (2011). Working toward a stronger conceptualization of scientific explanation for science education. *Science Education, 95*(4), 639-669.

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., Kidron, A., & Linn, M. C. (2019). Guiding collaborative revision of science explanations. International Journal of Computer-Supported Collaborative Learning, 14(3), 291–324. https://doi.org/10.1007/s11412-019-09298-y

Gerard, L., Linn, M. C., & Berkeley, U. C. (2022). Computer-based guidance to support students' revision of their science explanations. Computers & Education, 176, 104351.

Gerard, L., Linn, M. C., & Madhok, J. (2016). Examining the Impacts of Annotation and Automated Guidance on Essay Revision and Science Learning In Looi, C. K., Polman, J. L., Cress, U., and Reimann, P. (Eds.). Transforming Learning, Empowering Learners: The International Conference of the Learning Sciences (ICLS) 2016, Volume 1. Singapore: International Society of the Learning Sciences.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81-112.

Lee, H. S., Pallant, A., Pryputniewicz, S., Lord, T., Mulholland, M., & Liu, O. L. (2019). Automated text scoring and real-time adjustable feedback: Supporting revision of scientific arguments involving uncertainty. *Science Education, 103*(3), 590-622.

Lee, H. S., Gweon, G. H., Lord, T., Paessel, N., Pallant, A., & Pryputniewicz, S. (2021). Machine learning-enabled automated feedback: supporting students' revision of scientific arguments based on data drawn from simulation. *Journal of Science Education and Technology, 30*(2), 168-192.

Linn, M. C. (2006). *The knowledge integration perspective on learning and instruction.* Cambridge University Press.

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

Rivard, L.O.P. (1994), A review of writing to learn in science: Implications for practice and research. J. Res. Sci. Roscoe, R. D., & McNamara, D. S. (2013). Writing Pal: Feasibility of an intelligent writing strategy tutor in the high school classroom. Journal of Educational Psychology, 105(4), 1010.Teach., 31: 969-983. https://doi.org/10.1002/tea.3660310910

Roscoe, R. D., & McNamara, D. S. (2013). Writing Pal: Feasibility of an intelligent writing strategy tutor in the high school classroom. Journal of Educational Psychology, 105(4), 1010.

Shute, V. J. (2008). Focus on formative feedback. Review of educational research, 78(1), 153-189.

Singh, P., Gnesdilow, D., Cang, X., Baker, S., Goss, W., Kim, C., ... & Puntambekar, S. (2022, July). Design of real-time scaffolding of middle school science writing using automated techniques. In *International Society for the Learning Sciences Conference*.

Stemler, S. (2001). An overview of content analysis. *Practical Assessment, Research & Evaluation*, *7*(17), 137-146.

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.

Vygotsky, L. S., & Cole, M. (1978). Mind in society: Development of higher psychological processes. Harvard university press

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.


## Acknowledgments

We thank the students and the teacher who participated in this study. The research reported in this study is supported by a grant from the National Science Foundation #2010483 to the last author.