

Connecting Student Feedback Understanding to Gains in Argumentative Writing and Revision

Diane Litman¹, Zhexiong Liu¹, Tiawnwen Li¹, Lindsay Clare Matsumura¹,
Richard Correnti¹, Elaine Wang²

¹University of Pittsburgh, ²RAND Corporation

{dlitman, tianwen.li, lclare, rcorrent}@pitt.edu, zhexiong@cs.pitt.edu, ewang@rand.org

Abstract

While automated writing evaluation (AWE) systems typically assess essay quality, we develop a process-oriented approach assessing revision alignment with feedback. Classroom deployments show that student perceptions of AWE feedback correlate with writing improvement. Furthermore, feedback designed to build revision knowledge can enhance student outcomes, shifting focus from product to process.

1 Introduction

Through argumentative writing, students make sense of ideas - often from source texts - by identifying evidence and connecting evidence to claims. Recognizing the importance of academic writing, automated writing evaluation (AWE) systems have been developed to assess writing skills and provide formative feedback to guide revision (Burstein et al., 2020; Shermis and Wilson, 2024).

However, when students respond to AWE (and human) feedback, revisions often result in no improvement or even a decline in quality (MacArthur et al., 2008; Graham et al., 2015; Wang et al., 2020). One reason for this might be that students lack the necessary knowledge and skills for effective revision in response to feedback (Roscoe et al., 2013). Revision is a complex skill, requiring knowledge of evaluative criteria for effective writing within a genre. The ability to revise has in fact been shown to distinguish novice and expert writers (Bereiter and Scardamalia, 1987; Graham et al., 1995; Midgett et al., 2008; De La Paz and McCutchen, 2011; MacArthur, 2016). Feedback on argumentative writing content might be of limited utility without guidance and support to also develop revision knowledge.

To explore the utility of providing revision in addition to content feedback in AWE systems for improving student outcomes, we conduct an experiment comparing two AWE systems focused on stu-

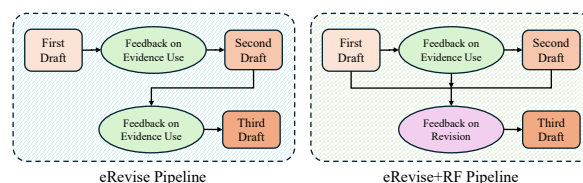


Figure 1: eRevise versus eRevise+RF system usage pipelines when students work on three essay drafts.

dents' text-based evidence usage in argumentative writing. The first system (eRevise) provides formative feedback on evidence usage from a source text, while eRevise+RF (RF stands for "revision feedback") provides feedback on both evidence usage and on revision, e.g., whether a student successfully implements the AWE system's prior feedback on evidence usage. Figure 1 compares eRevise vs. eRevise+RF in a three draft scenario. Outcomes relating to self-reported student perceptions of AWE feedback, as well as to system assessments of student writing and revision, are collected.

For example, assume the AWE evidence feedback on a student's first draft was to add more text-based evidence, and in response, the student instead corrected grammatical errors. While eRevise's feedback on the second draft would again be limited to evidence usage (e.g., "Add more pieces of evidence from the article to make your argument more convincing"), eRevise+RF's feedback would also include knowledge about revision (e.g., "When you revised your essay, I noticed that you edited your writing. Revising is different from editing to correct spelling or grammar errors. When writers revise their essays, they usually add new information from a text to support their argument.") Such feedback on revisions made in response to evidence feedback could potentially help students develop knowledge of successful revision in response to feedback.

We hypothesize that students' understanding of

AWE feedback is positively predictive of their revision success, and that as a result, feedback on the content of their argument writing (evidence use) in addition to revision feedback will improve students' knowledge of revision, leading toward more successful revision in response to feedback and ultimately more successful writing. In other words, we believe that feedback that provides information to students about their revision process will build students' knowledge of revision and ability to successfully implement the feedback they received on the quality of their use of evidence. Our experiment thus aims to answer the following two questions:

RQ1: Do survey responses regarding AWE feedback and perceived essay improvement relate to objective improvements in evidence usage in writing and to revision in response to feedback?

RQ2: Does providing AWE feedback on the revision process (eRevise+RF) in addition to providing feedback on evidence usage (eRevise) improve student outcomes?

Our results show that student self-reported perceptions of AWE feedback do indeed positively correlate with writing and revision outcomes (RQ1), and that eRevise+RF – designed to increase revision knowledge and skill – can improve student outcomes compared to eRevise (RQ2).

2 Related Work

Prior studies of AWE at the college (Roscoe et al., 2017) and high school (Roscoe et al., 2018) level have found that while student perceptions of AWE feedback are generally positive and that perceptions formed over time relate to revising, relationships between immediate perceptions of feedback on a particular essay and how students revise were minimal. Our work also examines immediate perception, but focuses on a younger student population.

AWE systems typically provide formative feedback on the quality of writing content (e.g. text-based evidence usage) rather than on the revisions undertaken after AWE feedback, even when multiple drafts are submitted (Foltz and Rosenstein, 2017; Mayfield and Butler, 2018). Revision as a skill has thus not been the target of AWE feedback¹, partially due to a lack of NLP technologies

for assessing alignment between AWE feedback and revision success (Burstein et al., 2020).

While initial NLP research on revision assessment evaluated revision in general terms (e.g., strength (Tan and Lee, 2014) or overall improvement (Afrin and Litman, 2018)), recent work has started to evaluate the extent to which revisions are aligned with feedback previously provided (Afrin and Litman, 2023; Liu et al., 2023; Li et al., 2026). We used such algorithms to develop a pilot version of eRevise+RF, informally observing that evidence usage in student essays improved in ways that were consistent with the feedback they received (Liu et al., 2025). However, our pilot study did not include an experimental control condition, and only student writing outcomes in terms of text-based evidence usage were collected. In the current study, we use eRevise to serve as an experimental control condition. We further expanded student outcomes to include: 1) measures of revision success in responding to AWE feedback; 2) student perceptions of AWE feedback and of essay improvement in evidence use.

3 Methods

3.1 The Response-to-Text Assessment

The Response-to-Text Assessment (RTA) (Correnti et al., 2013) is a validated argumentative writing task for assessing students' ability to reason using text-based evidence from a source article and to successfully use such evidence to support a claim. Currently there are two forms of the RTA: MVP and SPACE. MVP is based on an article about the United Nations Millennium Villages Project to fight poverty, while SPACE is based on an article about the benefits and costs of space exploration:

- MVP writing task: The author described how the quality of life can be improved by the Millennium Villages project in Sauri, Kenya. Based on the article, did the author convince you that “winning the fight against poverty is achievable in our lifetime”? Explain why or why not with 3 to 4 examples from the text to support your answer.
- SPACE writing task: Consider the reasons given in the article for why we should and should not fund space exploration. Did the

¹Alternative methods for student development beyond feedback messages have been to enhance AWE with other learn-

ing opportunities, e.g., providing writing strategy tutorials or game-based practice (Wilson and Czik, 2016; Roscoe et al., 2013), to help students better understand AWE feedback.

author convince you that “space exploration leads to long-term benefits” that justify the cost? Give reasons for your answer. Support your reasons with 3 to 4 pieces of evidence from the text.

To administer the RTA, teachers first read the source text aloud, asking questions and defining vocabulary to ensure that students understand the article before beginning to write.

3.2 The eRevise and eRevise+RF AWE Systems

eRevise and eRevise+RF (Liu et al., 2025) are AWE systems providing formative feedback to students taking the RTA. They are designed for classroom usage over multiple days to allow students to draft an essay, then revise it up to two times² in response to system feedback (Figure 1).

3.2.1 eRevise

Like most AWE systems, eRevise provides *formative feedback on the quality of the most recently submitted essay draft*. To do so, NLP is first used to assess students’ text-based evidence usage in a single essay draft in terms of two features operationalizing parts of the validated³ RTA holistic Evidence Score (ES) rubric used for human scoring: evidence breadth (EB) and evidence depth (ED)⁴. The NLP algorithms we developed for computing these features in eRevise and eRevise+RF (full details in Liu et al. (2025)) are summarized as follows. After an essay is embedded into a semantic space with Glove (Pennington et al., 2014), if a sliding essay window passes a similarity threshold to an RTA text’s evidence topics (breadth) or specific examples (depth), the window is counted as containing text-based evidence. The Evidence Breadth (EB) and Depth (ED) features are both integer-valued, with the MVP and SPACE evidence topics and examples taken from Rahimi et al. (2017).⁵ The holistic Evidence Score (ES) for an essay is an integer

²The original version of eRevise (Zhang et al., 2019) supported only one revision.

³RTA rubric construct validity was assessed by relating human-coded rubric scores to literacy instruction and student achievement (Correnti et al., 2013).

⁴These NLP assessment features were introduced to build (Rahimi et al., 2017) and validate (Correnti et al., 2020) an automated RTA holistic evidence scoring system. EB and ED were called NPE (numbers of pieces of evidence) and SPC (specificity of evidence), respectively, in our prior work.

⁵As a simplified example, the MVP article includes evidence across multiple topics, including hospitals (e.g., “The Yala Sub-District Hospital has medicine, free of charge, for all of the most common diseases.”) However, the following

ranging from 1 (low) to 4 (high), and is predicted with a BERT (Devlin et al., 2019) model trained on previously collected essays human-scored using the RTA evidence rubric.

Once the NLP features are computed, they are then used to select the most appropriate system formative feedback from a set of pre-defined messages. The messages reflect research on effective feedback and conceptual frameworks for text evidence use (Wang et al., 2018), and were created by content experts after an analysis of previously collected and scored RTA essays. eRevise currently selects from these expert-crafted messages rather than using LLM-generated feedback to increase the likelihood of offering more specific and actionable suggestions (Behzad et al., 2024). Table 5 in the Appendix shows the set of evidence feedback messages, and how they are selected based on the NLP evidence breadth and depth features.

3.2.2 eRevise+RF

eRevise+RF not only provides evidence feedback after the first draft, but after submission of the second draft, also provides *formative feedback on the quality of the revisions between the first and second drafts*. The NLP algorithms, which assess whether revisions align with the feedback provided and improve the essay with respect to evidence use (full details in Liu et al. (2025)), are summarized as follows. After drafts are aligned at the sentence-level using Bertalign (Liu and Zhu, 2022), a BERT-based binary classifier labels each essay revision (nonidentical aligned sentences) as surface (meaning-preserving) or content (meaning-altering). A ChatGPT-based binary classifier predicts whether each content revision is an evidence or reasoning revision (e.g., addresses the evidence or reasoning feedback in Table 5). A revision success (RS) binary classifier (a DistilRoBERTa (Sanh et al., 2019) model based on Liu et al. (2023)) predicts whether each sentence-level evidence or reasoning revision successfully aligned with the AWE feedback. Because all feedback is content-oriented, surface revisions are unsuccessful by default.

Once the NLP features are computed, an expert-crafted formative feedback message on revision is selected using the revision classifier predictions and the EB and ED values discussed earlier. Table 6

sentence from a student essay is not semantically similar to this text-based evidence, so the EB and ED counts would not be incremented: “Maybe starting with recycling and taking shorter showers, but no really short that you don’t get clean.”

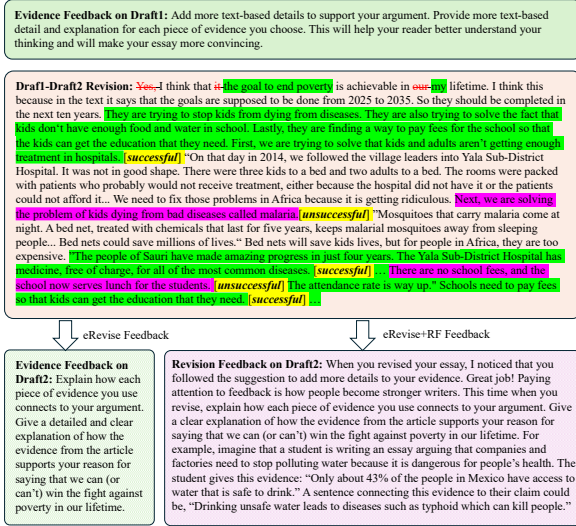


Figure 2: eRevise versus eRevise+RF formative feedback guiding student revision for third version of essay. The highlighted green and purple indicate added texts, and the yellow indicates attached revision labels, either successful or unsuccessful. The crossed red texts are deletions.

in the Appendix shows the revision feedback messages, and how they are selected based on the NLP assessments. The messages either help a student try again after unsuccessful revision, or help advance the student's essay after successful revision (e.g., rows 3-4 versus row 5 in Table 6). Figure 2 compares the eRevise versus eRevise+RF feedback after the same second draft of a student essay.

3.3 Classroom Experiment and Sample

In Fall 2025, we used eRevise and eRevise+RF to provide formative feedback to 5th-7th grade U.S. English Language Arts students taking both forms of the RTA. To examine the influence of the revision feedback, we implemented the following research design. In an initial three-day deployment (using the MVP writing task) with 12 teachers spanning six school districts within one geographical region, each teacher's students were randomly assigned to one of two system conditions - eRevise or eRevise+RF. For the second three-day deployment (using the SPACE writing task), the assignment of students to system remained the same. To address our RQs, we analyzed only data from students who completed all three drafts of each writing task and all surveys.

3.4 Outcome Measures

3.4.1 Writing Improvement (ΔEB , ΔED , ΔES) and Revision Success (#RS)

To evaluate improvement in *text-based evidence usage*, for every draft, we used the systems' NLP algorithms to compute EB, ED, and ES (§3.2), then measured improvement after revision by comparing these values across different essay drafts (ΔEB , ΔED , ΔES). To evaluate *successful revision* in implementing feedback, we used the revision success (RS) binary classifier (§3.2) to count the number of revised sentences between drafts that successfully implemented feedback (#RS).

To examine the reliability of the NLP algorithms for computing EB and ED, trained annotators (human-human ICCs of $\geq .85$) coded the essays of 20% randomly selected students from our sample (§3.3). Essay drafts were randomly ordered before human coding. When compared to the human annotations, NLP achieved EB/ED ICCs of 0.85/.90 for MVP, and 0.61/0.88 for SPACE. The two other NLP algorithms for writing outcomes were previously shown to have a Quadratic Weighted Kappa of 0.89/0.94 for ES for MVP/SPACE, and an F1 of 0.7 for RS across MVP and SPACE (Liu et al., 2025).

3.4.2 Student Perceptions of System Feedback ($F_{understand}$, $F_{knowledge}$, F_{use}) and of Essay Improvement ($E_{improve}$)

Student surveys were administered within the system immediately after each essay draft was submitted. Table 1 shows the Likert-scale questions administered after both drafts 2 and 3, and that are relevant to our research questions.

4 Results

Table 2 presents descriptives after both rounds of revision, for each RTA form. For the survey questions about AWE feedback (rows 1-3), all values are positive (above each response scale's 1.5 midpoint), indicating the perceived usability and utility of AWE feedback. This is similar to prior findings with high school and college students showing favorable immediate perceptions of AWE feedback (Roscoe et al., 2017, 2018). The positive values for essay improvement (row 4, all above scale midpoint of 1) also show that students believe their essays improve after each revision. The writing outcomes (rows 5-7) for both RTAs show improvements (positive Δs) from both first to second and second to third drafts,

Measure	Survey Question	Response Scale
$F_{understand}$	How well did you understand the feedback you received?	Not at all (0), A little bit (1), Mostly (2), Completely (3)
$F_{knowledge}$	How well did you know how to use the feedback to improve your essay?	Not at all (0), A little bit (1), Mostly (2), Completely (3)
F_{use}	How much of the feedback did you use when you revised your essay?	Not at all (0), A little (1), A lot (2), All (3)
$E_{improve}$	How much better do you think your revised essay is compared to your first draft?	Not at all (0), A little bit (1), A lot (2)

Table 1: Student survey questions on system feedback (rows 1-3) and perceived essay improvement (row 4), administered in the system after both rounds of essay revision (i.e., after submission of drafts 2 and 3).

	MVP (N=354)		SPACE (N=306)	
	1 to 2	2 to 3	1 to 2	2 to 3
$F_{understand}$	2.12	2.16	2.24	2.18
$F_{knowledge}$	1.97	2.01	2.08	2.05
F_{use}	1.96	2.01	2.05	2.04
$E_{improve}$	1.52	1.44	1.48	1.44
ΔEB	0.64	0.21	1.06	0.47
ΔED	3.54	1.35	2.99	1.50
ΔES	0.82	0.36	0.71	0.35
#RS	6.86	3.75	5.74	3.75

Table 2: Descriptives for first (drafts 1-2) and second (drafts 2-3) rounds of student essay revision, for both RTA forms. Rows 1-4: mean student response for survey questions on feedback (scale 0-3) and essay improvement (scale 0-2). Rows 5-8: means of automatically computed measures of text-based evidence usage improvement, and of number of revised sentences successfully implementing system feedback on prior draft.

with larger Δ s after the first round of revision. A similar pattern between revision rounds is observed for #RS.⁶ To dig deeper into these observations, we turn to our RQs.

4.1 RQ1: Relating Survey to Writing/Revision

As discussed in §1, a lack of substantial improvement after students revise in response to feedback might be because many students are weak in understanding and applying the feedback. In Table 3, we provide evidence for this possibility by examining whether positive student responses to questions about system feedback correlate with objective measures of revision success and writing improvement. With one exception, the number of successful revisions (#RS) does indeed significantly and positively correlate with each of the three survey response measures about feedback; these rela-

tionships are consistent across both revision cycle and RTA form, although the correlations would conventionally be interpreted as weak (0.10-0.39) at best (Schober et al., 2018). Second, there are also many significant (again weak) positive correlations between the three writing improvement measures (the deltas) and the three feedback survey response measures. These positive relationships between immediate perceptions of feedback and writing/revision improvement for our 5th-7th graders are different than prior findings at the high school and college level (Roscoe et al., 2017, 2018). Finally, the significant, weak positive correlations in almost all cells of the two $E_{improve}$ rows indicate that students are able to self-assess both their writing and revision success.

4.2 RQ2: Effect of eRevise versus eRevise+RF

The theory of change behind eRevise+RF was that feedback on success in addressing draft 1 feedback would increase understanding and implementation of draft 2 feedback and draft 3 writing outcomes. To test this, we followed the experimental design in §3.3. Table 4 presents the between-condition results. Since eRevise and eRevise+RF vary in the feedback that is seen when students are working on draft 3, we only consider draft 3 survey outcomes and overall writing improvements from draft 1 to 3.⁷

First, all outcomes were higher for students in the eRevise+RF condition, with one exception (ΔEB for MVP). Second, many of these differences were statistically different ($p < .05$) with

⁶Essays typically contain 10-20 sentences.

⁷We report draft 1 to 3 deltas in Table 4 as total growth across system condition is the practical quantity of interest, e.g., do eRevise+RF students end up with better essays than eRevise students? Draft 1 to 3 is also the delta for $E_{improve}$ (Table 1). The draft 2 - 3 deltas, which provide a more isolated test of the revision feedback manipulation, can be found in Table 7 in the Appendix and yield similar observations.

	MVP Draft 1 to 2				SPACE Draft 1 to 2			
	ΔEB	ΔED	ΔES	#RS	ΔEB	ΔED	ΔES	#RS
$F_{understand}$	.219*	.181*	.064	.116*	.084	.085	.089	.076
$F_{knowledge}$	.191*	.167*	.124*	.155*	.067	.118*	.134*	.148*
F_{use}	.149*	.092	.105*	.148*	.068	.129*	.162*	.212*
$E_{improve}$	.258*	.346*	.158*	.261*	.186*	.241*	.278*	.320*
	MVP Draft 2 to 3				SPACE Draft 2 to 3			
	ΔEB	ΔED	ΔES	#RS	ΔEB	ΔED	ΔES	#RS
$F_{understand}$	.019	.098	.062	.196*	-.036	-.012	.114*	.164*
$F_{knowledge}$	.066	.096	.061	.179*	.025	.071	.118*	.117*
F_{use}	.035	.117*	.014	.184*	.054	.065	.048	.160*
$E_{improve}$	.092	.128*	.068	.150*	.068	.178*	.235*	.211*

Table 3: Spearman correlation coefficients between student survey responses (rows) and system assessments (columns), across RTA forms and revision phases. *: $p < .05$. N=354 students for MVP, 306 for SPACE.

After Draft 3	RTA MVP				RTA SPACE			
	eRevise N=188	eRevise+RF N=166	Sig	Eta Sq	eRevise N=152	eRevise+RF N=154	Sig	Eta Sq
$F_{understand}$	2.06	2.28	.008	.020	2.11	2.26	.078	.010
$F_{knowledge}$	2.00	2.01	.880	.000	1.97	2.14	.053	.012
F_{use}	1.93	2.11	.015	.017	1.94	2.13	.029	.016
$E_{improve}$	1.42	1.46	.445	.002	1.36	1.52	.006	.024
ΔEB (draft 1 to 3)	0.85	0.84	.901	.000	1.37	1.69	.047	.013
ΔED (draft 1 to 3)	4.69	5.11	.442	.002	3.95	5.02	.044	.013
ΔES (draft 1 to 3)	1.14	1.23	.445	.002	1.05	1.07	.875	.000

Table 4: ANOVA analyses of system condition effects. Top: mean response after draft 3 submission, for survey questions about feedback. Middle: perceived draft 1 to 3 improvement. Bottom: NLP draft 1 to draft 3 improvements on text-based evidence usage. Best condition results *italicized*. Significant condition differences ($p < .05$) **bolded**.

small effect sizes (Eta Squared $> .01$). With respect to perceptions of AWE feedback, for both MVP and SPACE, students in the eRevise+RF condition reported using more of the feedback they received than students in the eRevise condition (F_{use}). This suggests that the revision feedback messages of eRevise+RF were perceived as being used more frequently than the evidence feedback messages of eRevise, and were thus more likely to be implemented. For MVP, $F_{understand}$ was also significantly higher for eRevise+RF students, suggesting that revision feedback was clearer to students than evidence feedback alone. With respect to writing outcomes, for SPACE, the improvements from original to final draft, measured using either self-report ($E_{improve}$) or NLP-assessed evidence breadth (ΔEB) or depth (ΔED), were significantly higher for students using eRevise+RF.

In sum, our results suggest that eRevise+RF improves students' perceptions of the feedback they receive and is associated with greater improve-

ments in the quality of text-based evidence in writing. Understanding why the MVP writing results are weaker is left for future work, e.g., it might reflect RTA form difficulty or ordering effects (MVP before SPACE (§3.3)).

5 Conclusion

We hypothesized that enhancing eRevise to provide feedback about the quality of students' evidence use in conjunction with guidance for implementing eRevise feedback had potential to support revision skill development. Classroom deployments of eRevise versus eRevise+RF show that positive student Likert-scale responses to questions about system feedback correlate with measures of both revision success (the number of revised sentences between drafts that successfully implemented feedback) and writing improvement (change in breadth, depth, and holistic usage of evidence across drafts). In addition, the deployments suggest that the process-oriented eRevise+RF with feedback designed to in-

crease revision knowledge and skill both improves students' perceptions of the feedback they receive and is associated with greater improvements in the quality of revision and writing outcomes.

In future work, we would like to expand our survey questions about AWE feedback, motivated by similar research with older students. We also plan to incorporate LLM-based methods that we have concurrently been developing to both address computational limitations of our existing NLP algorithms (e.g., the need for large amounts of training data), as well as to support new eRevise+RF feedback functionalities requiring generative capabilities such as dynamically creating feedback messages that are more personalized to student essays or that instantiate different theoretical frameworks.

Limitations

Our results are based on a relatively small essay corpus, collected from classrooms within one geographical region. Also, although 587 students for MVP and 565 students for SPACE submitted at least one draft, many were absent from class on some days of data collection, or were present but submitted essays but no surveys. The current evaluation considers only the students for whom we have complete survey and essay data (356 students for MVP, 306 for SPACE). We note that we are currently not correcting for multiple comparisons in these first analyses of only a subset of our students, and thus there is the potential for false positives in the presented results. Also, since only 224 of these students are in the intersection of these MVP and SPACE students (i.e., have complete survey and essay data across both RTA forms), we have not yet examined longitudinal effects from the MVP to the SPACE deployment. Using statistical techniques such as imputation to mitigate risks from omitting missing data is future work.

Our evidence and revision outcomes are based on NLP rather than human annotations. Although we have manually annotated subsets of our data to demonstrate the reliability as well as the validity of using NLP-based outcomes in both our prior work and the current study (e.g., last paragraph of §3.4.1), the potential for noisy results is increased.

Ethics Statement

eRevise and eRevise+RF were deployed under protocols approved by an institutional review board (IRB). Student essay and survey data are not pub-

licly available due to IRB, to ensure young student privacy.

Acknowledgments

This research was supported by National Science Foundation Award #2202347 and by its REU supplement and by a supplement made in response to NSF DCL 24-093. We also acknowledge the National Artificial Intelligence Research Resource (NAIRR) Pilot and Cloudbank allocations for contributing to our research results. The opinions expressed are those of the authors and do not represent the views of the institutes.

References

- Tazin Afrin and Diane Litman. 2018. Annotation and classification of sentence-level revision improvement. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 240–246, New Orleans, Louisiana.
- Tazin Afrin and Diane Litman. 2023. [Predicting desirable revisions of evidence and reasoning in argumentative writing](#). In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2550–2561, Dubrovnik, Croatia. Association for Computational Linguistics.
- Shabnam Behzad, Omid Kashefi, and Swapna Soma-sundaran. 2024. [Assessing online writing feedback resources: Generative AI vs. good samaritans](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1638–1644, Torino, Italia. ELRA and ICCL.
- Carl Bereiter and Marlene Scardamalia. 1987. An attainable version of high literacy: Approaches to teaching higher-order skills in reading and writing. *Curriculum inquiry*, 17(1):9–30.
- Jill Burstein, Brian Riordan, and Daniel McCaffrey. 2020. Expanding automated writing evaluation. In *Handbook of automated scoring*, pages 329–346. Chapman and Hall/CRC.
- Richard Correnti, Lindsay Clare Matsumura, Laura Hamilton, and Elaine Wang. 2013. Assessing students' skills at writing analytically in response to texts. *The Elementary School Journal*, 114(2):142–177.
- Richard Correnti, Lindsay Clare Matsumura, Elaine Wang, Diane Litman, Zahra Rahimi, and Zahid Kisa. 2020. Automated scoring of students' use of text evidence in writing. *Reading Research Quarterly*, 55(3):493–520.
- Susan De La Paz and Deborah McCutchen. 2011. Learning to write. *Handbook of research on learning and instruction*, pages 46–68.

- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Peter W Foltz and Mark Rosenstein. 2017. Data mining large-scale formative writing. *Handbook of learning analytics*, 199:199–210.
- Steve Graham, Karen R Harris, and Tanya Santangelo. 2015. Research-based writing practices and the common core: Meta-analysis and meta-synthesis. *The Elementary School Journal*, 115(4):498–522.
- Steve Graham, Charles MacArthur, and Shirley Schwartz. 1995. Effects of goal setting and procedural facilitation on the revising behavior and writing performance of students with writing and learning problems. *Journal of Educational Psychology*, 87(2):230.
- Tianwen Li, Lindsay Clare Matsumura, Elaine Wang, Diane Litman, Zhexiong Liu, Richard Correnti, and Michelle Hong. 2026. [Ai-assisted evaluation of revision patterns in young students' argument writing](#). *Frontiers in Education*, Volume 11 - 2026.
- Lei Liu and Min Zhu. 2022. Bertalign: Improved word embedding-based sentence alignment for chinese–english parallel corpora of literary texts. *Digital Scholarship in the Humanities*.
- Zhexiong Liu, Diane Litman, Elaine Wang, Lindsay Matsumura, and Richard Correnti. 2023. [Predicting the quality of revisions in argumentative writing](#). In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 275–287, Toronto, Canada. Association for Computational Linguistics.
- Zhexiong Liu, Diane Litman, Elaine L Wang, Tianwen Li, Mason Gobat, Lindsay Clare Matsumura, and Richard Correnti. 2025. [eRevise+RF: A writing evaluation system for assessing student essay revisions and providing formative feedback](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 173–190, Albuquerque, New Mexico. Association for Computational Linguistics.
- Charles A MacArthur. 2016. Instruction in evaluation and revision.
- Charles A MacArthur, Steve Graham, and Jill Fitzgerald. 2008. *Handbook of writing research*. Guilford Press.
- Elijah Mayfield and Stephanie Butler. 2018. Districtwide implementations outperform isolated use of automated feedback in high school writing. In *International Conference of the Learning Sciences*, volume 2128.
- Ekaterina Midgette, Priti Haria, and Charles MacArthur. 2008. The effects of content and audience awareness goals for revision on the persuasive essays of fifth- and eighth-grade students. *Reading and Writing*, 21(1):131–151.
- Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.
- Zahra Rahimi, Diane Litman, Richard Correnti, Elaine Wang, and Lindsay Clare Matsumura. 2017. Assessing students' use of evidence and organization in response-to-text writing: Using natural language processing for rubric-based automated scoring. *International Journal of Artificial Intelligence in Education*, 27(4):694–728.
- Rod D Roscoe, Laura K Allen, Adam C Johnson, and Danielle S McNamara. 2018. Automated writing instruction and feedback: Instructional mode, attitudes, and revising. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 62, pages 2089–2093. SAGE Publications Sage CA: Los Angeles, CA.
- Rod D. Roscoe, Erica L. Snow, and Danielle S. McNamara. 2013. Feedback and revising in an intelligent tutoring system for writing strategies. In *Artificial Intelligence in Education*, pages 259–268, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Rod D Roscoe, Joshua Wilson, Adam C Johnson, and Christopher R Mayra. 2017. Presentation, expectations, and experience: Sources of student perceptions of automated writing evaluation. *Computers in Human Behavior*, 70:207–221.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. [Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter](#). In *EMC2: Efficient Methods for Deep Learning Workshop at NeurIPS 2019*.
- Patrick Schober, Christa Boer, and Lothar A Schwarte. 2018. Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, 126(5):1763–1768.
- Mark D Shermis and Joshua Wilson. 2024. *The Routledge international handbook of automated essay evaluation*. Taylor & Francis.
- Chenhao Tan and Lillian Lee. 2014. A corpus of sentence-level revisions in academic writing: A step towards understanding statement strength in communication. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2: Short Papers, pages 403–408, Baltimore, MD, USA.
- E. Wang, L. C. Matsumura, and R. Correnti. 2018. Student writing accepted as high-quality responses to analytic text-based writing tasks. *The Elementary School Journal*, 118(3):357–383.

Elaine Lin Wang, Lindsay Clare Matsumura, Richard Correnti, Diane Litman, Haoran Zhang, Emily Howe, Ahmed Magooda, and Rafael Quintana. 2020. erevis (ing): Students’ revision of text evidence use in an automated writing evaluation system. *Assessing Writing*, 44:100449.

Joshua Wilson and Amanda Czik. 2016. Automated essay evaluation software in english language arts classrooms: Effects on teacher feedback, student motivation, and writing quality. *Computers & Education*, 100:94–109.

Haoran Zhang, Ahmed Magooda, Diane Litman, Richard Correnti, Elaine Wang, LC Matsumura, Emily Howe, and Rafael Quintana. 2019. erevise: Using natural language processing to provide formative feedback on text evidence usage in student writing. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9619–9625.

A Appendix

NLP Essay Assessment	Feedback Description	Evidence Feedback Message
Evidence Breadth (EB) $\leq \alpha$	Needs more evidence	•Add more pieces of evidence from the article to make your argument more convincing. •Reread the highlighted portions of the article to choose more evidence that helps your side of the argument about whether or not we can win the fight against poverty in our lifetime.
Evidence Breadth (EB) $> \alpha$, Evidence Depth (ED) $\leq \beta$	Needs more specific details	•Add more text-based details to support your argument. •Provide more text-based detail and explanation for each piece of evidence you choose. This will help your reader better understand your thinking and will make your essay more convincing.
Evidence Breadth (EB) $> \alpha$, Evidence Depth (ED) $> \beta$	Needs more reasoning	•Explain how each piece of evidence you use connects to your argument. •Give a detailed and clear explanation of how the evidence from the article supports your reason for saying that we can (or can't) win the fight against poverty in our lifetime.

Table 5: Formative feedback messages (column 3) on evidence (after drafts 1 and 2 in eRevise, only after draft 1 in eRevise+RF), and how they are selected using the NLP features for assessing evidence usage (column 1). The threshold α for EB is set to half of the topics discussed in the relevant RTA source article, and β for ED is set to half of the example categories. Note that the messages shown in column 3 are for the MVP form of the RTA, and are slightly modified for SPACE. For example, for SPACE, the 2nd bullets in the first and third rows of the table are “Choose more evidence from the GREEN highlighted text if you support funding Space Exploration, from the PINK highlighted text if you think we should not fund Space Exploration.” and “Give a detailed and clear explanation of how the evidence from the article supports your reason for saying that we should (or should not) fund space exploration.”, respectively.

NLP Essay and Revision Assessments	Feedback Description	Revision Feedback Message
No content and no surface revisions, or only sentence deletions	No attempt	- When you revised your essay, I noticed very few changes between your first and second draft. When writers revise their essays, they usually add new information to support their argument. - Reread the highlighted portions of the article to choose more evidence that helps your side of the argument about whether or not we can win the fight against poverty in our lifetime.
No content revisions	Surface revision	- When you revised your essay, I noticed that you edited your writing. Revising is different from editing to correct spelling or grammar errors. When writers revise their essays, they usually add new information from a text to support their argument. - Reread the highlighted portions of the article to choose more evidence that helps your side of the argument about whether or not we can win the fight against poverty in our lifetime.
No Evidence Breadth (EB) change; Added sentences contain evidence topic words already used	Repeats evidence	- When you revised your essay, I noticed that you added in evidence that was very similar to the evidence you had included before. - Also, add more detail for each piece of evidence you use. For example, imagine that a student is writing an essay about water pollution, the sentence "People don't have enough water" would not be detailed enough. It would be better to write, "Only about 43% of the people in Mexico have access to water that is safe to drink." - Reread the highlighted portions of the article to choose more evidence that helps your side of the argument about whether or not we can win the fight against poverty in our lifetime.
No Evidence Breadth (EB) change; Added sentences contain no evidence topic words	Added evidence but not text-based	- Reread your essay to make sure that your evidence is from the article. This time when you revise, make sure to add more evidence to support your claim. - Also, add more detail for each piece of evidence you use. For example, imagine that a student is writing an essay about water pollution, the sentence "People don't have enough water" would not be detailed enough. It would be better to write, "Only about 43% of the people in Mexico have access to water that is safe to drink." - Reread the highlighted portions of the article to choose more evidence that helps your side of the argument about whether or not we can win the fight against poverty in our lifetime.
Evidence Breadth (EB) change (new EB > old EB)	Added evidence successfully	- When you revised your essay, I noticed that you followed the suggestion to add more evidence. Great job! Paying attention to feedback is how people become stronger writers. - This time when you revise, add more details for each piece of evidence you use. For example, imagine that a student is writing an essay about water pollution, the sentence "People don't have enough water" would not be detailed enough. It would be better to write, "Only about 43% of the people in Mexico have access to water that is safe to drink." - In addition, explain how each piece of evidence you use connects to your argument. Give a clear explanation of how the evidence from the article supports your reason for saying that we can (or can't) win the fight against poverty in our lifetime.
Evidence Breadth (EB) change (new EB > old EB), but Evidence Depth (ED) changes $\leq$ to γ	Added evidence but vague or general	- When you revised your essays, I noticed that you added more evidence from the text. A good writer also adds details to the evidence they use to make the essay more convincing. - This time when you revise your essay, add more details for each piece of evidence you use. For example, if a student is writing an essay about water pollution, the sentence "People don't have enough water" would not be detailed enough. It would be better to write "Only about 43% of the people in Mexico have access to water that is safe to drink." - In addition, explain how each piece of evidence you use connects to your argument. Give a clear explanation of how the evidence from the article supports your reason for saying that we can (or can't) win the fight against poverty in our lifetime.
Evidence Depth (ED) changes > γ (new ED > old ED)	Added specific details successfully	- When you revised your essay, I noticed that you followed the suggestion to add more details to your evidence. Great job! Paying attention to feedback is how people become stronger writers. - This time when you revise, explain how each piece of evidence you use connects to your argument. Give a clear explanation of how the evidence from the article supports your reason for saying that we can (or can't) win the fight against poverty in our lifetime. - For example, imagine that a student is writing an essay arguing that companies and factories need to stop polluting water because it is dangerous for people's health. The student gives this evidence: "Only about 43% of the people in Mexico have access to water that is safe to drink." A sentence connecting this evidence to their claim could be, "Drinking unsafe water leads to diseases such as typhoid which can kill people."
No reasoning revision	Successful evidence but no reasoning attempt	- When you revised your essay, I noticed that you added evidence or details from the article. A good writer also explains why they use each piece of evidence and how this evidence supports their argument. - This time when you revise, explain how each piece of evidence you use connects to your argument. For example, imagine that a student is writing an essay arguing that companies and factories need to stop polluting water because it is dangerous for people's health. The student gives this evidence "Only about 43% of the people in Mexico have access to water that is safe to drink." A sentence connecting this evidence to their claim could be "Drinking unsafe water leads to diseases such as typhoid which can kill people." - In addition, include a counterclaim. This means bringing up a good point that is against the side you are arguing for, but then explaining why the evidence that supports your position is stronger.
Unsuccessful reasoning revision	Successful evidence but unsuccessful reasoning	- When you revised your essay, I noticed that you did not fully explain how your evidence connects to your claim. - This time when you revise, make sure that you explain how each piece of evidence you use connects to your argument. For example, imagine that a student is writing an essay arguing that companies and factories need to stop polluting water because it is dangerous for people's health. The student gives this evidence: "Only about 43% of the people in Mexico have access to water that is safe to drink." A sentence connecting this evidence to their claim could be, "Drinking unsafe water leads to diseases such as typhoid which can kill people." - In addition, include a counterclaim. This means bringing up a good point that is against the side you are arguing for, but then explaining why the evidence that supports your position is stronger.
Successful reasoning revision	Successful evidence and successful reasoning	- When you revised your essay, I noticed that you followed the suggestion to explain your evidence and how it connects to your argument. Great job! Paying attention to feedback is how people become stronger writers. - This time when you revise your essay, include a counterclaim. This means bringing up a point that is against the side you are arguing for, but then explaining how that may actually support your side. - For example, imagine that a student is writing an essay arguing that there needs to be laws to make companies and factories pay for polluting water. A counterclaim that acknowledges and addresses an opposing view might be "Some people argue that making companies pay will make things like clothes and cars more expensive. But paying doctors to heal diseases caused by water pollution is expensive also."- Using a counterclaim makes your argument stronger. It shows that you have considered the opposite side but have stronger reasons for supporting your position.

Table 6: eRevise+RF revision feedback messages after draft 2 (column 3), and how they are selected using NLP (column 1).

	RTA MVP				RTA SPACE			
	eRevise N=188	eRevise+RF N=166	Sig	Eta Sq	eRevise N=152	eRevise+RF N=154	Sig	Eta Sq
Δ EB (draft 2 to 3)	<i>0.24</i>	0.17	.215	.004	0.38	<i>0.56</i>	.051	.012
Δ ED (draft 2 to 3)	<i>1.49</i>	1.19	.282	.003	1.30	<i>1.70</i>	.175	.006
Δ ES (draft 2 to 3)	0.31	<i>0.42</i>	.211	.004	0.33	<i>0.38</i>	.568	.001

Table 7: ANOVA analyses of system condition effects on draft 2 to draft 3 improvements on text-based evidence usage. Best condition results *italicized*. As with the draft 1 to 3 results in Table 4, the draft 2 to 3 effect sizes are small at best, and the benefits of eRevise+RF over eRevise are primarily for RTA SPACE.