

EXPLORING THE ROLE OF FEEDBACK IN AI AUTOMATED ESSAY SCORING

Rianto^{1*}, Widya Syafitri^{2*}

^{1,2} State Islamic University of Sjech M. Djamil Djambek Bukittinggi, Indonesia

Email Correspondence: rian.anto002@gmail.com

©2024 by the authors. Submitted for open access publication under the terms and conditions of the Creative Commons Attribution-ShareAlike 4.0 International License-(CC-BY-SA) (<https://creativecommons.org/licenses/by-sa/4.0/>)

DOI : <http://dx.doi.org/10.30983/icmie/>

ABSTRACT

Automated Essay Scoring (AES) systems have gained popularity in education for their potential to efficiently and objectively grade student writing. However, concerns have been raised about their accuracy and the lack of feedback they provide to students. In this research, the writers aim to explore the role of feedback in AI Automated Essay Scoring. First, writers reviewed the literature on AES systems, their effectiveness and the role of feedback in student writing development. The writers conducted a quantitative study using tests and surveys to gather data from students. The test was used to gather the students' writing before and after giving them feedback provided by AI, while the survey focused on students' perceptions of AES feedback. The findings revealed that the provision of feedback through AES significantly improved the quality of student writing generally. Furthermore, the survey results indicated that students perceived the AES feedback as helpful and informative, with the majority of students reporting that the feedback helped them identify areas for improvement and develop their writing skills.

Keywords: Feedback, AES, Writing, AI

INTRODUCTION

Automated Essay Scoring (AES) systems have become increasingly popular in education settings, with the promise of efficient and objective grading of student writing. However, concerns have been raised about the accuracy of these systems and feedback provided to students. Feedback is an essential component in the process of writing development, as it helps students improve their writing skills and understand their strengths and weaknesses. As such, the role of feedback in AI Automated Essay Scoring has become a critical area of research.

Many studies have focused on the use of Automated Writing Evaluation (AWE) tools for providing automated written corrective feedback in classrooms to improve student writing (Cotos, 2014; Koltovskaia, 2020; Ranalli, 2013). Popular AWE tools like Criterion and MY Access! have been evaluated for their effectiveness in classroom settings (Chen & Cheng, 2008; Dikli & Bleyle, 2014; Z. Li et al., 2014). However, there is some hesitation in using these AWE tools as they may not align with the writing skills valued in the classroom (Condon, 2013). Nevertheless, studies have found that using AWE tools like Criterion can lead to increased revisions and improved accuracy, ultimately positively impacting the quality of student writing (J. Li et al., 2015; Stevenson & Phakiti, 2014).

However, since the launch of GPT by Open-AI, the feedback provided by AI has improved significantly.

For example:

In learning English, they are four skills that should be mastered By students; they are listening, speaking, reading, and writing. Those skills need language component to support them. One of the language components is vocabulary. The vocabulary is one of language component which is very important to support each skill in learning English. Because without vocabulary students will be difficult to comprehend and understand each skill in learning English. Thornbury says, "without vocabulary nothing can be conveyed. If the students' lack the vocabulary students have a problem in learning English such as in speaking they difficult to express their idea because lack of vocabulary. In writing, the student will difficult to express their idea in writing text because lack the vocabulary. Furthermore, in reading student difficult to understand comprehend the text because they lack vocabulary. So student must have ability in mastered vocabulary to support all of skill in English.

Figure 1. Input (Students' Writing)

The text above is one of the paragraphs of the students' thesis. This was then inputted to the AI with a certain prompt. Furthermore, the result was like:

I would give this writing a score of 2 out of 5. Here's my assessment and feedback:

Strengths:

The writer has attempted to organize their ideas in a logical sequence, starting with the importance of mastering four skills in learning English and then highlighting the crucial role of vocabulary in supporting these skills.

The writer has included a quote from Thornbury to add credibility to their argument.

Weaknesses:

The writing lacks clarity and coherence, with several grammatical errors, awkward sentence structures, and unclear transitions between sentences.

The writer repeats themselves excessively, using phrases like "in learning English" multiple times, which makes the text tedious to read.

The vocabulary is limited, and the writer relies heavily on simple sentences, which makes the text lack variety and depth.

The text lacks concrete examples, illustrations, or anecdotes to make the points more engaging and relatable.

Figure 2. Output (Holistic Score Given by GPT 4.0)

Not only the holistic score, feedback of strength, and weaknesses of the writing, the AI also provided the suggestion for the writing improvement. It can be seen such as below:

Suggestions for improvement:

1. Revise the sentence structures to improve clarity and coherence. Consider using a mix of simple, compound, and complex sentences to create a more dynamic rhythm.
2. Vary your vocabulary to convey nuanced meanings and avoid repetition. For instance, instead of using "difficult" repeatedly, try "challenging," "struggling," or "facing obstacles."
3. Use transitions to connect your ideas more smoothly. For example, you could use "Moreover," "Furthermore," or "In addition" to link your sentences more effectively.
4. Provide concrete examples or illustrations to make your points more relatable and engaging. This could include scenarios, anecdotes, or hypothetical situations that demonstrate the importance of vocabulary in learning English.
5. Edit your work carefully to eliminate grammatical errors, punctuation mistakes, and typos. This will help you present your ideas more professionally and confidently.

Figure 3. Output (AI's Suggestion for improvement)

By giving the AI a simple prompt as an instruction to give a holistic score of a student's writing, it provided the score with elaborate explanation as well as the feedback to improve the writing. This could provide new insight for teachers to grade and give feedback for students and the feedback would also be useful for the students.

Regarding the teachers and students' perception toward the feedback that were given by AES, there are some previous researches that studied about the topic. Four studies have examined teachers' perceptions towards the use of automated feedback systems in the classroom, but have found conflicting results. Grimes and Warschauer (2008) found that most teachers were positive towards using automated feedback and believed it saved them time and provided specific feedback for students. Chen and Cheng (2008) found that teachers were not confident in the accuracy of the scores and supplemented the automated feedback with their own. Grimes and Warschauer (2010) and Klobucar et al. (2013) both reported that teachers were not convinced of the system's accuracy and had little desire to use it in their classrooms. These findings suggest that while automated feedback can be helpful, it may not be as accurate as human feedback. It is important for teachers to provide a balance of both types of feedback and teach students how to interpret and use the automated feedback critically.

Studies have shown mixed perceptions towards the use of automated feedback systems by students. Some studies (Grimes & Warschauer, 2008) reported that students were motivated to improve their writing skills with the instant scores and feedback provided by the system. However, others (Chen & Cheng, 2008; Grimes & Warschauer, 2010; Klobucar et al., 2013) found that students had negative perceptions, citing issues such as abstract and vague feedback, overemphasis on surface level issues, and lack of

human understanding. The preference for longer essays by automated systems was also questioned as it may not accurately reflect writing ability (Chandrasegaran, 2013).

However, these studies were from pre-GPT era, there was no research about the teachers or students' perception about feedback provided by GPT. This makes it important for this research to be conducted as it can provide valuable insights on the potential use of GPT as an automated feedback system in the classroom. Moreover, GPT is a relatively new artificial intelligence technology that has not been extensively studied in the context of feedback provision. Its ability to generate human-like text and adapt to different writing styles may address some of the concerns raised in previous studies about automated feedback systems. For example, GPT may provide more specific and personalized feedback compared to traditional automated systems, which may help students to improve their writing skills. Additionally, the use of GPT may save teachers time in providing feedback, allowing them to focus on other aspects of teaching and learning.

Automated Essay Scoring (AES)

The reference list should be arranged alphabetically following the guidelines of the Publication Manual of the American Psychological Association (7th ed.). For example:

Automated Essay Scoring (AES) is the use of computer programs and algorithms to evaluate and score written essays. The software analyzes various aspects of the essay, such as grammar, vocabulary, sentence structure, and content, and generates a score based on predefined criteria or rubrics. AES is often used in educational settings, particularly for large-scale assessments and standardized tests, to provide rapid and consistent evaluation of student writing. It is also used in online learning platforms to provide feedback on student writing and improve their writing skills.

Automated Essay Scoring (AES) is a technology that uses computer programs to evaluate and score written essays based on predefined criteria. The use of AES has both advantages and disadvantages, but it is generally recommended because it can avoid the risks associated with human scoring. This technology has been continuously studied and improved upon since the 1960s, with Page (1967) being the pioneer behind the Project Essay Grade (PEG).

Early AES systems, such as PEG, focused mainly on surface structures and were criticized for neglecting content-related features. However, with advancements in Natural Language Processing (NLP), more precise AES systems were developed in the 1990s. These systems, including e-rater, used a combination of statistical and rule-based methods to analyze not only surface structures but also deeper linguistic features. As a result, e-rater has been officially used in high-stakes tests such as the GRE and TOEFL.

Moreover, AES systems often have additional automated feedback engines, leading to research on Automated Writing Evaluation (AWE) for classroom use. Critics argue that these tools measure different writing constructs than those valued in the

classroom, but others have found that they can help learners increase revisions and accuracy, leading to better-text quality (J. Li et al., 2015; Stevenson & Phakiti, 2014). For instance, e-rater's AWE tool, Criterion, has been researched for its effectiveness in the classroom (e.g., Chen & Cheng, 2008; Dikli & Bleyle, 2014; Z. Li et al., 2014).

In conclusion, Automated Essay Scoring (AES) utilizes computer technology to evaluate and score written essays based on predefined criteria. Its use has grown since the 1960s, with advancements in NLP leading to more accurate systems such as e-rater. Additionally, the feedback function in AES systems has been studied for its potential use in the classroom, with mixed opinions on its effectiveness.

Generative Pre-trained Transformer (GPT)

GPT (Generative Pre-trained Transformer) is a transformer-based language model that has been developed by OpenAI. It is trained on a massive amount of text data and has been fine-tuned for various natural language tasks such as text generation, language translation, and question-answering. Unlike BERT, which is a bidirectional model, GPT is a unidirectional model, meaning that it only reads the text from start to end. GPT has been known for its exceptional performance in language generation tasks. It has been released in three versions: GPT-1, GPT-2, and GPT-3, with each version having a larger size and more capabilities. GPT-3 is currently the largest with 175 billion parameters. While GPT has been used for various natural language tasks, it has not been utilized in AES. However, recent advancements in GPT technology, such as the release of ChatGPT, have made it more accessible and user-friendly, leading to speculations that it could be used in AES applications in the future (Essel, 2023).

GPT and BERT are both transformer-based language models, but they have different architecture and training approaches. BERT is bidirectional and trained by filling in missing words in sentences, while GPT is unidirectional and trained on a large corpus of text for natural language generation tasks. This difference in training and architecture makes BERT more suitable for natural language understanding tasks, while GPT is better at language generation tasks. In AES, where both understanding and generation of language are important, BERT has been utilized due to its effectiveness in sentiment analysis and text classification, which are similar to AES tasks. However, recent research (Mayer et al., 2023) has shown that GPT can also perform complex tasks such as classification of e-mail replies as polite or impolite with comparable accuracy to human ratings.

Despite its limitations, such as not being able to undergo task-specific fine-tuning and mainly functioning as a language generation model, GPT's prompt-based learning approach makes it a more user-friendly and accessible option for non-experts in AES. This approach eliminates the need for programming and has been shown to attain accuracy levels similar to human ratings in classification tasks (Mayer et al., 2023). The recent release of ChatGPT, a web-based platform that is free to use, has also enhanced the accessibility of GPT technology and has led to speculations of its potential use in AES applications (Essel, 2023). Overall, GPT has shown great potential in various NLP tasks

and with advancements in technology and research, it could potentially be utilized in AES in the future.

RESEARCH METHOD

The writers conducted a quantitative study using a test and survey to gather data from students. The test was used to gather the students' writing before and after getting the AI feedback, while the survey focused on students' perceptions of AES feedback. The test, which consisted of a standardized writing task, was administered to a sample of 65 students which were chosen randomly from the English Education Study Program in UIN Bukittinggi. The survey, which comprised 15 Likert-scale items, was designed to elicit students' attitudes and experiences with AES feedback. The data collected from the test and survey were analyzed using descriptive statistics and inferential statistics. The test results were analyzed to identify any significant differences in the students' writing quality before and after receiving the AI feedback. The results of the study were then compared to existing literature to provide a comprehensive understanding of the effectiveness of AES feedback in improving students' writing skills

RESULTS AND DISCUSSION

The results of the study revealed a significant improvement in the quality of students' writing after receiving feedback from the AI Automated Essay Scoring system. The mean score of the students' writing before receiving feedback was 65.21, which increased to 74.15 after receiving feedback. The paired-samples t-test revealed a statistically significant difference between the two mean scores ($t = 5.646$, $p < 0.01$). This suggests that the provision of feedback through AES had a positive impact on the quality of students' writing.

Paired Samples Test

		t	df	Sig. (2-tailed)
Pair 1	Pre - Post	5.646	24	0.000

A significant difference between student writing before and after the use of AI feedback can be observed in two ways:

1) By comparing the calculated t-value with the table t-value:

Since $5.646 > 2.064$, there is a difference between the writing before and after the implementation of AI feedback.

2) By comparing the significance value with alpha:

Since $0.000... < 0.05$, there is a difference between the writing before and after the implementation of AI feedback.

In terms of students' perceptions of AES feedback, the survey results indicated that the majority of students (85%) found the feedback to be helpful and informative. Most students (80%) reported that the feedback helped them identify areas for improvement in their writing, and 75% of students indicated that the feedback helped them develop their writing skills. Furthermore, 70% of students reported that they

were more motivated to improve their writing skills after receiving feedback from the AES system.

However, a smaller proportion of students expressed some reservations about the feedback, citing concerns about the accuracy and clarity of the comments provided. Despite this, the overall sentiment among students was positive, with many appreciating the timely and personalized feedback that the AES system provided.

CONCLUSIONS AND SUGGESTIONS

The findings of this study suggest that the integration of AI Automated Essay Scoring systems in educational settings can have a positive impact on the quality of students' writing. The statistically significant improvement in students' writing scores and the positive perceptions of students towards AES feedback indicate that this technology has the potential to support writing instruction and improve student learning outcomes.

Based on these findings, it is recommended that educators and policymakers consider the integration of AI-powered feedback systems in educational settings to support student writing development. Additionally, further research is needed to explore the long-term effects of AES feedback on student writing development and to address the concerns raised by students regarding the accuracy and clarity of the feedback provided. By doing so, we can harness the potential of AI technology to enhance teaching and learning practices, ultimately leading to better student outcomes.

REFERENCES

- Applebee, A., & Langer, J. (2011). A snapshot of writing instruction in secondary and high schools. *English Journal*, 100, 14–27.
- Bahri, D., Tay, Y., Zheng, C., Brunk, C., Metzler, D., & Tomkins, A. (2021, March). Generative models are unsupervised predictors of page quality: A colossal-scale study. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining* (pp. 301-309).
- Bang, H. J. (2013). Reliability of national writing project's analytic writing continuum assessment system. *Journal of Writing Assessment*, 6(1), 1-12.
- Beseiso, M., Alzubi, O. A., & Rashaideh, H. (2021). A novel automated essay scoring approach for reliable higher educational assessments. *Journal of Computing in Higher Education*, 33, 727-746.
- Brown, G. T. L. (2009). The reliability of essay scores: The necessity of rubrics and moderation. *Tertiary Assessment and Higher Education Student Outcomes: Policy, Practice and Research*, 40-48.
- Chen, L., Zaharia, M., & Zou, J. (2023, July 18). How is ChatGPT's behavior changing over time? Retrieved from <https://arxiv.org/pdf/2307.09009.pdf>
- Cohen, Y., Levi, E., & Ben-Simon, A. (2018). Validating human and automated scoring of essays against “True” scores. *Applied Measurement in Education*, 31(3), 241-250.
- Crossley, S., Baffour, P., Yu, T., Franklin, A., Benner, M., & Boser, U. (2023). A large-scale corpus for assessing written argumentation: PERSUADE 2.0. doi: 10.1016/j.asw.2023.100667

- Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, 25(2), 155–185.
- Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication*, 32(4), 365-387.
- Graham, S. (2018). A revised writer(s)-within-community model of writing. *Educational Psychologist*, 53(4), 258-279.
- Grimes, D., & Warschauer, M. (2010). Utility in a fallible tool: A multi-site case study of automated writing evaluation. *Journal of Technology, Learning, and Assessment*, 8(6), 1-44.
- Ifenthaler, D. (2022). Automated essay scoring systems. In *Handbook of open, distance and digital education* (pp. 1-15). Springer Nature Singapore.
- Ifenthaler, D., & Dikli, S. (2015). Automated scoring of essays. In J. M. Spector (Ed.), *The SAGE encyclopedia of educational technology* (Vol. 1, pp. 64–68). Sage.
- Klein, J., & Taub, D. (2005). The effect of variations in handwriting and print on evaluation of student essays. *Assessing Writing*, 10(2), 134-148.
- Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 363-374.
- Lawrence, J. F., Galloway, E. P., Yim, S., & Lin, A. (2013). Learning to write in secondary school? *Journal of Adolescent & Adult Literacy*, 57(2), 151-161.
- Liu, Y., Moosavi, N. S., & Lin, C. (2023). LLMs as narcissistic evaluators: When ego inflates evaluation scores. *arXiv preprint arXiv:2311.09766*.
- Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2), 100050.
- Monte-Sano, C. (2010). Disciplinary literacy in history: An exploration of the historical nature of adolescents' writing. *The Journal of the Learning Sciences*, 19(4), 539-568.
- Monte-Sano, C. (2012). What makes a good history essay? Assessing historical aspects of argumentative writing. *Social Education*, 76(6), 294-298.
- Monte-Sano, C., & De La Paz, S. (2012). Using writing tasks to elicit adolescents' historical reasoning. *Journal of Literacy Research*, 44(3), 273-299.
- National Center for Education Statistics. (2012). *The Nation's Report Card: Writing 2011* (NCES 2012–470). Institute of Education Sciences, U.S. Department of Education, Washington, D.C.
- National Writing Project. (2010). *The Analytic Writing Continuum: A comprehensive writing assessment system*. University of California, Berkeley; Berkeley, CA: National Writing Project.
- Northwest Regional Educational Laboratory. (2011). 6+1 Trait Writing. Retrieved from <http://educationnorthwest.org/traits>
- Olson, C. B., Woodworth, K., Arshan, N., Black, R., Chung, H., D'Aoust, C., & Dewar, T. (2020). The pathway to academic success: Scaling up a text-based analytical writing intervention for Latinos and English learners in secondary school. *Journal of Educational Psychology*, 112(4), 701–717.
- Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, 88(2), 413–428.
- Shermis, M. D., & Burstein, J. (Eds.). (2013). *Handbook of automated essay evaluation: Current applications and new directions*. Routledge.

- Stephen, T. C., Gierl, M. C., & King, S. (2021). Automated essay scoring (AES) of constructed responses in nursing examinations: An evaluation. *Nurse Education in Practice*, 54, 103085.
- Tate, T. P., Warschauer, M., & Abedi, J. (2016). The effects of prior computer use on computer-based writing: The 2011 NAEP writing assessment. *Computers & Education*, 101, 115-131.
- Uto, M., Aomi, I., Tsutsumi, E., & Ueno, M. (2023). Integration of prediction scores from various automated essay scoring models using item response theory. *IEEE Transactions on Learning Technologies*.
- Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing*, 6(2), 145-178.