

The Effect of AI-Powered Automated Writing Assessment on EFL Students' Writing Performance

Irfan Maulana Marhudi^{1*}, Sri Wahyuni²

^{1,2}Pascasarjana UIN Syekh Wasil Kediri, East Java, Indonesia

*Corresponding author: maulanamarhudi@gmail.com

Abstract:

This study aims to evaluate the effectiveness of Grammarly as an Artificial Intelligence (AI)-powered Automated Writing Assessment (AWA) tool in enhancing the writing performance of EFL students. A quasi-experimental design was employed, involving 61 ninth-grade students at MTs Al Islah, divided into an experimental group (n=32) receiving Grammarly-assisted feedback and a control group (n=29) receiving traditional teacher feedback. Data were collected through pre-tests and post-tests using an argumentative essay task, which were then analysed using descriptive statistics and t-tests via SPSS 25. The results indicated a significant improvement in both groups; however, the experimental group achieved a notably higher mean score (85.50) compared to the control group (75.00). Statistical analysis revealed a significant difference between the two groups (Sig. 2-tailed = 0.000), suggesting that AI-powered feedback serves as a more potent catalyst for writing improvement than conventional methods. The findings imply that Grammarly effectively reduces the cognitive load on students by providing instant linguistic corrections, thereby allowing for better focus on content and organization. This study recommends the integration of AI tools as a formative assessment in EFL pedagogy to foster self-regulated learning in the digital era.

Keywords: AI-Powered Assessment; EFL Writing; Grammarly; Automated Writing Assessment (AWA)

Abstrak:

Penelitian ini bertujuan untuk mengevaluasi efektivitas Grammarly sebagai alat Penilaian Menulis Otomatis (Automated Writing Assessment/AWA) berbasis Kecerdasan Buatan (Artificial Intelligence/AI) dalam meningkatkan kinerja menulis siswa EFL. Desain kuasi-eksperimental digunakan, melibatkan 61 siswa kelas sembilan di MTs Al Islah, dibagi menjadi kelompok eksperimen (n=32) yang menerima umpan balik dengan bantuan Grammarly dan kelompok kontrol (n=29) yang menerima umpan balik guru tradisional. Data dikumpulkan melalui pre-test dan post-test menggunakan tugas esai argumentatif, yang kemudian dianalisis menggunakan statistik deskriptif dan uji-t melalui SPSS 25. Hasil menunjukkan peningkatan yang signifikan pada kedua kelompok; namun, kelompok eksperimen mencapai skor rata-rata yang jauh lebih tinggi (85,50) dibandingkan dengan kelompok kontrol (75,00). Analisis statistik mengungkapkan perbedaan yang signifikan antara kedua kelompok (Sig. 2-tailed = 0,000), menunjukkan bahwa umpan balik berbasis AI berfungsi sebagai katalis yang lebih ampuh untuk peningkatan menulis daripada metode konvensional. Temuan ini menunjukkan bahwa Grammarly secara efektif mengurangi beban kognitif siswa dengan memberikan koreksi linguistik instan, sehingga memungkinkan fokus yang lebih baik pada konten dan organisasi. Studi ini merekomendasikan integrasi alat AI sebagai penilaian formatif dalam pedagogi EFL untuk mendorong pembelajaran mandiri di era digital.

Kata kunci: Penilaian Berbasis AI; Penulisan EFL; Grammarly; Penilaian Penulisan Otomatis (AWA)



INTRODUCTION

In the era of globalization, English as a Foreign Language (EFL) proficiency has become indispensable for academic and professional success, particularly in writing skills that demand accuracy, coherence, and organization. However, EFL students, especially in non-English speaking countries like Indonesia, often struggle with writing due to limited exposure to native-like input, fossilized errors, and insufficient feedback on complex grammatical structures and rhetorical patterns (Miranty et al., 2023). Traditional classroom-based writing assessment, reliant on teacher feedback, faces challenges such as time constraints, subjective evaluations, and scalability issues, leading to delayed or inconsistent improvements in students' performance.

The advent of Artificial Intelligence (AI)-powered Automated Writing Assessment (AWA) tools has revolutionized EFL writing pedagogy by providing instant, scalable, and data-driven feedback. One of the technology-based tools employed in the development of EFL/ ESL learners' language skills is artificial intelligence (AI). Indeed, there are various AI-based applications that have been used to develop the writing skills of EFL/ESL learners. These applications include, among others, KAKU, Quill Bot, Word tune, and Grammarly. Tools like Grammarly exemplify this shift, leveraging natural language processing algorithms to evaluate grammar, vocabulary, style, and overall clarity in real-time. Unlike conventional methods, AWA systems offer formative assessments that empower learners to revise iteratively, fostering self-regulated learning and enhancing writing proficiency (Alangari, 2025). This technology aligns with modern educational demands for personalized and efficient instruction.

Writing in EFL contexts represents a multifaceted productive skill encompassing linguistic accuracy, rhetorical organization, lexical richness, and

ideational content. Miranty et al., (2023) It requires not only mechanical proficiency but also genre awareness and critical thinking, which EFL learners frequently underperform due to L1 interference and lack of authentic practice. While, Marghany (2023) clarified that FL writing is not a simple task which depends on just presenting ideas; it includes multiple processes to accomplish the final form of the writing task. Thus, EFL learners are required to plan a main structure in which they should accurately synthesize sentences and paragraphs. Effective writing instruction thus demands timely, specific feedback to bridge these gaps, positioning AI tools as transformative agents in skill development.

Recent studies underscore the potential of AI in EFL contexts. For instance, Haziz et al. (2025) demonstrated that AI tools like ChatGPT significantly improved writing skills among Indonesian EFL senior high school students by enhancing content generation and structure. Similarly, Marghany (2023) found that AI-based instruction boosted essay writing skills in EFL higher education through targeted feedback on mechanics and organization. In Indonesia, Miranty et al. (2023) explored automated writing evaluation tools, revealing positive impacts on undergraduate EFL students' accuracy and fluency.

Despite these advancements, perceptions of AI-assisted writing remain mixed among users. Rofikah et al. (2025) reported that Indonesian students viewed AI tools positively for academic writing support, yet highlighted concerns over over-reliance and creativity suppression. Bessadat and Korichi (2025) noted Algerian EFL teachers' favourable perceptions of AI for formative assessment but emphasized the need for empirical evidence on actual performance gains. Marzuki et al. (2023) further indicated that EFL teachers observed improvements in content and organization via AI tools, though integration challenges persisted.

The present study addresses this gap by investigating Grammarly's effect on EFL students' writing performance through a quasi-experimental approach. By comparing an experimental group using Grammarly with a control group receiving traditional feedback, this research quantifies improvements in holistic scores (content, organization, grammar, vocabulary) and identifies mediating factors like motivation and error reduction patterns. Grounded in the socio-cognitive framework of writing development, it contributes to EFL pedagogy by providing

evidence-based recommendations for AI integration. To this end, we posed the following critical research questions: (1) To what extent does the use of Grammarly as an AI-powered AWA tool improve EFL students' overall writing performance compared to traditional teacher feedback? (2) What specific writing components (grammar, vocabulary, organization, content) show the most significant gains from Grammarly integration? (3) How do students' perceptions of Grammarly influence their writing performance outcomes?

The urgency in answering these research questions stems from two primary reasons. Firstly, quantifying the specific gains in components like grammar and organization provides an empirical blueprint for educators to integrate AI-powered AWA tools like Grammarly more effectively. This ensures that technology serves as a precise catalyst for improving EFL writing performance. Secondly, investigating students' perceptions is vital, as their attitudes toward AI influence their self-regulated learning and iterative revision habits. Understanding whether students view these tools as beneficial or as a risk to creativity allows for better instructional strategies to ensure the potential of AI is fully realized. By answering these questions, this study hopes to pave the way for more evidence-based AI integration in EFL classrooms.

RESEARCH METHOD

This research uses quantitative method. According to Latief (2012) Using Quantitative method in the form of numbers and calculation. This design of this research is experimental. Experimental research assesses the effect of manipulated and controlled variable (as independent variable) to another variable (as dependent variable). The specific design of this research is quasi-experimental. It is because researchers cannot create groups artificially for experiments. This approach was chosen to evaluate the impact of the AI-assisted learning intervention on writing proficiency, as outlined by Mahmoudi (2017). Although some studies have explored the application of AI tools in EFL contexts, this research specifically addresses the need for the practical implementation of these tools in high school settings.

A total of 61 EFL students from class 9 at MTs Al Islah participated in this study, divided into two groups: an experimental group (n=32) receiving AI-powered feedback via Grammarly and a control group (n=29) receiving traditional teacher feedback. Participants were selected using purposive sampling to ensure homogeneity in terms of age (13-15 years), proficiency level (A2-B1 based on CEFR), and prior exposure to English writing tasks.

The research instrument consisted solely of a writing test in the form of an argumentative essay task (150-300 words) on EFL-relevant topics. Both tests consisted of a written essay on a specific topic, Jacobs et al., (1981) analytical scoring rubric was used to assess the writing skills due to its proven reliability on EFL writing skills (Chang et al., 2024). The rubric covering five components: content (30%), organization (20%), vocabulary (20%), grammar (20%), and mechanics (10%), with a total score of 100. Pre-test and post-test were administered to both groups to measure changes in writing ability before and after the intervention.

Data were collected across five meetings for each group during the second semester of 2026, with each session lasting 90 minutes in the computer lab at MTs Al Islah to ensure consistent access to technology. Meeting 1 involved administering the pre-test to establish baseline writing proficiency levels for all participants. In Meeting 2, both groups conducted initial writing practice, followed by assessment: the experimental group used Grammarly for automated feedback, while the control group received teacher feedback. Meeting 3 focused on a second writing task for both groups to apply feedback iteratively. Meeting 4 repeated the assessment phase, with Grammarly for the experimental group and teacher feedback for the control group. Finally, Meeting 5 concluded with the post-test to evaluate overall improvements.

The researcher uses quantitative forms to analyse the data. Writing scores from pre-test and post-test were analysed using SPSS 25. Paired t-tests measured pre-post differences within each group, while independent t-tests compared differences between groups. ANCOVA was used with pre-test scores as covariates to control for initial equivalence. Cohen's d effect sizes were calculated for practical significance, with a significance level of $p < 0.05$. Normality (Shapiro-Wilk) and homogeneity (Levene) assumptions were verified beforehand.

RESULTS AND DISCUSSION

Descriptive Patterns of Students' Achievement

This section presents the descriptive analysis of the research data collected. The analysed data comprise the pre-test and post-test scores from both the experimental and control groups, aimed at illustrating the achievement and score distribution of the students before and after the treatment was administered. A summary of the descriptive statistics for this data is presented in Table 3.1 below.

Table 1. Descriptive Statistics of Pre-test and Post-test Scores

	N	Minimum	Maximum	Mean	Std. Deviation
Pretest Experimental	32	62	75	68.16	2.908
Posttest Experimental	32	80	92	85.50	3.427
Pretest Control	29	55	78	64.76	5.932
Posttest Control	29	65	82	75.00	4.276
Valid N (listwise)	29				

Source: Primary data, processed by the researcher (2026)

As shown in Table 1, a notable increase in mean scores is evident for both groups. In the experimental group, the average score rose significantly from 68.16 in the pre-test to 85.50 in the post-test. Meanwhile, the control group also showed improvement, with the mean score increasing from 64.76 to 75.00.

Furthermore, both the minimum and maximum scores in both groups shifted upward following the treatment. The relatively low standard deviation values across all categories indicate that the data within each group were stable and did not deviate significantly from the mean. These descriptive results provide an initial indication that there is a difference in achievement between the experimental and control groups following the implementation of the learning intervention. The following inferential statistical tests will further examine whether this observed difference is statistically significant.

Assessment of Statistical Assumptions

Before conducting inferential statistical analyses, tests of statistical assumptions were performed to assess the suitability of the data for parametric testing. These included a normality test to examine the distribution of the pre-test and post-test scores and a homogeneity of variance test to assess the equality of variances between the experimental and control groups. The results are presented in Tables 2 and Figure 1, respectively.

Table 2. Results of the Normality Test

	Class	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
Results	1	.166	32	.025	.946	32	.115
	2	.098	32	.200*	.949	32	.135
	3	.130	29	.200*	.939	29	.096
	4	.179	29	.019	.929	29	.052

*. This is a lower bound of the true significance.

Source: Primary data, processed using SPSS (2026).

Based on the normality test results presented in Table 2, an analysis was conducted to determine whether the data follow a normal distribution, which is a prerequisite for using parametric statistics. Given that the sample size in this study is relatively small ($N < 50$), the interpretation of the data focuses on the significance values from the Shapiro-Wilk test. The statistical analysis reveals that all four groups achieved a significance level (Sig.) greater than the 0.05 threshold. Specifically, the significance values obtained were 0.135 for Experimental group and 0.052 for Control group. Since all these values exceed 0.05, it can be concluded that the data across all groups are normally distributed. These findings provide a solid statistical foundation to proceed with hypothesis testing using parametric statistical procedures.

Moreover, a parametric test, namely the paired samples t-test, was conducted to examine whether there is a significant difference between the pre-test

and post-test scores within each group. The results of the paired samples test are presented in Figure 1 below.

Figure 1. Results of the Homogeneity of Variance Test

		Paired Samples Test							
		Paired Differences							
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference		t	df	Sig. (2-tailed)
					Lower	Upper			
Pair 1	PretesEksperimental - PostesEsperimental	-17.344	4.597	.813	-19.001	-15.686	-21.341	31	.000
Pair 2	PretesControl - PostesControl	-10.241	6.457	1.199	-12.697	-7.785	-8.542	28	.000

Source: Primary data, processed using SPSS (2026).

Based on Table 3.3, it can be seen that the significance values (Sig. 2-tailed) for both the experimental and control groups are 0.000, which are lower than 0.05. This indicates that there are significant differences between the pre-test and post-test scores in both groups. However, the mean difference in the experimental group is higher than that of the control group, suggesting that the treatment had a greater effect on improving students' achievement in the experimental group compared to the control group.

Within-Group Improvement and Between-Group Differences

Following the assessment of statistical assumptions, inferential analyses were conducted to examine changes in students' achievement within each group and differences between the experimental and control groups. A paired-samples t-test was used to compare pre-test and post-test scores within each group, while an independent-samples t-test was employed to compare the post-test scores between the two groups. The results of these analyses are presented in Tables 3 and Figure 2, respectively. Descriptive statistics comparing the post-test scores are provided in Table 4 to further illustrate the differences in achievement between the groups.

Table 3. Results of the Paired-Samples t-Test

		Levene Statistic	df1	df2	Sig.
Result	Based on Mean	1.803	1	59	.185
	Based on Median	.916	1	59	.342

Based on Median and with adjusted df	.916	1	54.099	.343
Based on trimmed mean	1.646	1	59	.205

Source: Primary data, processed using SPSS (2026).

Based on Table 3, the significance value of the Levene's test is greater than 0.05 (Sig. > 0.05). This indicates that the variance of the data between the experimental and control groups is homogeneous. Therefore, the assumption of homogeneity is fulfilled, and further parametric statistical tests can be appropriately applied.

Since the assumptions of normality and homogeneity have been fulfilled, the independent samples t-test was conducted to determine whether there is a significant difference between the experimental and control groups after the treatment. The results of the independent samples t-test are presented in Figure 2 below.

Figure 2. Results of the Independent-Samples t-Test

Independent Samples Test										
		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Hasil	Equal variances assumed	1.803	.185	10.629	59	.000	10.500	.988	8.523	12.477
	Equal variances not assumed			10.513	53.658	.000	10.500	.999	8.497	12.503

Source: Primary data, processed using SPSS (2026).

Based on Figure 2, the significance value (Sig. 2-tailed) is 0.000, which is lower than 0.05. This indicates that there is a statistically significant difference between the experimental and control groups. Furthermore, the mean difference shows that the experimental group achieved higher scores than the control group. Therefore, it can be concluded that the treatment had a significant positive effect on students' achievement. To further support the results of the independent samples t-test, descriptive statistics were also analysed to compare the mean scores of the experimental and control groups. The summary of the group statistics is presented in Table 4 below.

Table 4. Comparison of Post-test Scores Between the Experimental and Control Groups

	Class	N	Mean	Std. Deviation	Std. Error Mean
Results	1	32	85.50	3.427	.606
	2	29	75.00	4.276	.794

Source: Primary data, processed using SPSS (2026).

Based on Table 3.6, it can be seen that the mean score of the experimental group (85.50) is higher than that of the control group (75.00). This difference indicates that students in the experimental group performed better than those in the control group. Therefore, these descriptive findings further strengthen the result that the treatment was effective in improving students' achievement.

Discussion

The statistical findings of this study provide empirical evidence that the integration of Grammarly as an AI-powered Automated Writing Assessment (AWA) tool significantly enhances EFL students' writing performance. As evidenced by the paired samples t-test, the experimental group achieved a substantial mean score increase from 68.16 to 85.50, outperforming the control group which rose from 64.76 to 75.00. This disparity confirms that while traditional teacher feedback remains valuable, the immediacy and scalability of AI-driven feedback provide a more intensive catalyst for improvement. These results align with Alangari (2025), who emphasizes that AWA tools empower learners to engage in iterative revisions, thereby fostering self-regulated learning habits that are often difficult to achieve through delayed conventional feedback.

The significant achievement in the experimental group also mirrors the findings of Haziz et al. (2025) and Marghany (2023), who identified that AI tools effectively bridge the gap in linguistic accuracy and structural organization among Indonesian EFL learners. By providing real-time corrections on mechanics and grammar, Grammarly likely reduced the cognitive load on students at MTs Al Islah, allowing them to allocate more mental resources toward higher-order tasks such as

content development and rhetorical coherence. Furthermore, the homogeneity and normality of the data, as shown in Table 3.2 and Table 3.4, suggest that the benefits of this AI intervention were consistent across the student population, reinforcing the reliability of AWA tools in secondary education settings.

The disparity in mean differences between the two groups highlights a critical pedagogical shift in managing common EFL writing challenges, such as fossilized errors and L1 interference. While the control group improved, the experimental group's superior performance indicates that the formative nature of AI feedback serves as a more transformative agent in skill development. This technology aligns with modern educational demands for personalized instruction, as it allows each student to address specific grammatical and lexical weaknesses at their own pace. This finding supports the socio-cognitive framework of writing development, where the interaction between the digital tool and the learner facilitates a deeper internalization of language patterns through instant reinforcement.

CONCLUSION

Based on the research findings and discussion, it can be concluded that the integration of Grammarly as an AI-powered AWA tool significantly improves the writing performance of EFL students compared to traditional teacher feedback. The significant increase in the experimental group's mean scores from 68.16 to 85.50 demonstrates that immediate and data-driven feedback facilitates more effective iterative revisions than delayed manual assessments. Statistical evidence from the independent samples t-test confirms that this improvement is not only statistically significant but also practically superior in enhancing student achievement.

Furthermore, the implementation of Grammarly serves as a transformative agent by bridging the gap in linguistic accuracy and structural organization. By providing real-time reinforcements, the tool allows students to address specific grammatical and mechanical weaknesses at their own pace, aligning with the demands for personalized and efficient instruction. This research successfully validates that AI-assisted learning interventions are a decisive factor in superior writing outcomes within secondary education settings. Consequently, educators are

encouraged to adopt AI tools like Grammarly as a permanent supplementary fixture in EFL pedagogy to facilitate deeper internalization of language patterns and support academic success in the era of globalization.

REFERENCES

- Alangari, T. S. (2025). The effect of AI-assisted learning on EFL writing proficiency: Quasi-experimental and cluster analysis. *Educational Process: International Journal*, 17, Article e2025345. <https://doi.org/10.22521/edupij.2025.17.345> edupij.com
- Al-Raimi, M., Mudhsh, B. A., Al-Yafaei, Y., & Al-Maashani, S. (2024). Utilizing artificial intelligence tools for improving writing skills: Exploring Omani EFL learners' perspectives. *Forum for Linguistic Studies*, 6(2), Article 1177. <https://doi.org/10.59400/fls.v6i2.1177> Forum for Linguistic Studies
- Bessadat, A. I., & Korichi, S. (2025). Perceptions of AI-based assessment for formative writing in EFL: Evidence from Algerian teachers. *Aleph*, 12(3), 59–71. <https://aleph.edinum.org/14943> Aleph
- Chang, Y.-T., Choe, A. T., Holden, D., & Isbell, D. R. (2024). Making each point count: Revising a local adaptation of the Jacobs et al.'s (1981) ESL composition profile rubric. *Language Testing*, 41(2), 443–455. <https://doi.org/10.1177/02655322231217979> journals.sagepub.com
- Haziz, M. Z., Agustina, Y., Arfah, H., & Nazri, M. A. (2025). The role of AI ChatGPT in improving writing skills of Indonesian EFL senior high school students. *Journal of English Language and Education*, 10(5), 385–390. <https://doi.org/10.31004/jele.v10i5.1411> jele.or.id
- Hyland, K., & Hyland, F. (2006). Feedback on second language students' writing. *Language Teaching*, 39(2), 83–101. <https://doi.org/10.1017/S0261444806003399> Cambridge Core
- Latief, M. A. (2012). *Research methods on language learning: An introduction*. UM Press. Digilib UB
- Mahmoudi, A. (2017). Effect of planning on Iranian intermediate EFL learners' mastery of writing skill. *Theory and Practice in Language Studies*, 7(3), 219–226. <https://doi.org/10.17507/tpls.0703.08> academypublication.com
- Marghany, M. M. (2023). Using artificial intelligence-based instruction to develop EFL higher education students' essay writing skills. *CDELT Occasional Papers in the Development of English Education*, 82(1), 219–240. <https://doi.org/10.21608/opde.2023.313623> opde.journals.ekb.eg
- Marzuki, Widiati, U., Rusdin, D., Darwin, & Indrawati, I. (2023). The impact of AI writing tools on the content and organization of students' writing: EFL teachers' perspective. *Cogent Education*, 10(2), Article 2236469.

<https://doi.org/10.1080/2331186X.2023.2236469> doi.org

- Miranty, D., Widiati, U., Cahyono, B. Y., & Sharif, T. I. S. T. (2023). Automated writing evaluation tools for Indonesian undergraduate English as a foreign language students' writing. *International Journal of Evaluation and Research in Education*, 12(3), 1705–1715. <https://doi.org/10.11591/ijere.v12i3.24958> ijere.iaescore.com
- Rofikah, U., Arafah, B., Nasmilah, & Harewaty. (2025). Indonesian students' perception of AI-assisted EFL academic writing. In M. Hasyim, M. A. Armin, & Y. Yusuf (Eds.), *Proceedings of the 5th International Conference on Linguistics and Cultural Studies 5 (ICLC-5 2024)* (pp. 410–419). Atlantis Press. https://doi.org/10.2991/978-2-38476-394-8_46 Atlantis Press
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, Eds.). Harvard University Press. <https://doi.org/10.2307/j.ctvjf9vz4>