

ASSESSING EFL STUDENTS' ESSAY WRITING WITH AUTOMATED WRITING EVALUATION: VALIDITY AND RELIABILITY EVIDENCE

Nadya Safitri Wulandari^{1*}, Rania Anindita², Fikri Maulana³, Arif Rahman Hidayat⁴

^{1,2,3,4}Universitas Muhammadiyah Malang, Indonesia

*Corresponding author: nanadsafitw12@gmail.com

Abstract - This study aimed to examine the validity and reliability of Automated Writing Evaluation (AWE) in assessing EFL students' essay writing. A quantitative validation design was employed involving 60 undergraduate students and 60 argumentative essays evaluated by two human raters and an AWE system using an analytic rubric covering content and task achievement, organization, vocabulary, grammatical accuracy, and mechanics. Human and automated scoring reliability was examined using the Intraclass Correlation Coefficient (ICC), while validity was investigated through Pearson correlation, agreement analysis, paired-samples t-test, and Bland–Altman analysis. The findings showed that the AWE system demonstrated excellent test–retest reliability (ICC = .95). The automated scores also showed a strong correlation with human ratings ($r = .87$) and good overall agreement (ICC = .86). No significant difference was found between the AWE and human scores ($p = .749$). The strongest agreement was observed for mechanics and grammatical accuracy, whereas the weakest agreement was found for content and task achievement. These findings indicate that AWE has considerable potential as a consistent and valid assessment tool for EFL essay writing; however, human judgment remains necessary, particularly for evaluating complex and interpretive dimensions of writing. The findings are in line with recent research indicating that automated essay scoring can achieve strong reliability while still presenting limitations in construct representation and dimension-level scoring.

Keywords: Automated Writing Evaluation; EFL Writing; Automated Essay Scoring; Validity; Reliability.

1. INTRODUCTION

Essay writing is an important component of English as a Foreign Language (EFL) education because it requires learners to demonstrate their ability to organize ideas, develop arguments, select appropriate language, and communicate meaning coherently. In higher education, written performance is also frequently used as evidence of students' academic language proficiency and their ability to respond to complex communicative tasks. However, evaluating a large number of essays manually can require substantial time and effort from teachers and assessors, while differences in rater interpretation, severity, and attention may introduce inconsistency into the assessment process. These challenges have increased interest in technology-supported approaches to writing assessment, particularly Automated Writing Evaluation (AWE), which uses computational procedures to analyze written texts and generate scores or feedback (Petchprasert, 2021; Guo et al., 2021; Nunes et al., 2022).

AWE has developed from systems primarily designed to identify grammatical and mechanical errors toward more sophisticated technologies capable of examining multiple dimensions of written performance. Earlier AWE applications provided automated feedback on grammar, spelling, vocabulary, and sentence-level features, whereas contemporary systems increasingly incorporate artificial intelligence and large language models (LLMs) to evaluate broader aspects of writing. This development has made automated assessment more attractive in EFL contexts because it can potentially provide rapid, scalable, and relatively

History:

Received : 14 August 2026

Revised : 27 August 2026

Accepted : 1 September 2026

Published : 9 September 2026

Publisher: Universitas PGRI Palembang

Licensed: This work is licensed under a Creative Commons Attribution 4.0 License



consistent evaluation across large numbers of student essays (Yao et al., 2021; McCarthy et al., 2022; Ngo et al., 2022; Fleckenstein et al., 2023).

Empirical evidence generally indicates that AWE can contribute positively to students' writing development, although its effects are not uniform across writing dimensions. A three-level meta-analysis by Ngo et al. (2022), for example, reported positive overall effects of AWE on EFL/ESL writing, with stronger benefits observed under particular instructional conditions. Similarly, Fleckenstein et al. (2023) found a medium overall effect of automated feedback on writing performance, while Nunes et al. (2022) emphasized that the effectiveness of AWE is influenced by instructional context, writing practice, control conditions, and teacher involvement. Research on EFL learners has also shown that students can successfully respond to automated corrective feedback, although the quality and accuracy of their revisions depend partly on the quality of the feedback itself (Guo et al., 2021). These findings suggest that AWE has considerable pedagogical potential, but its usefulness cannot be established solely through improvements in writing performance.

A particularly important issue concerns what automated systems actually measure when they assign scores to essays. Automated scoring may demonstrate high agreement with human ratings while still relying heavily on linguistic features that do not fully represent the broader construct of writing proficiency. Ferrara and Qunbar (2022) therefore argued that automated scores require evidence concerning construct relevance, construct representation, transparency, and fairness. Shermis (2022) similarly highlighted the difficulty of establishing validity when automated systems measure selected textual characteristics rather than the complete construct of good writing. Evidence from English learner assessments further indicates that differences between human and automated scoring may vary according to learner characteristics and testing conditions (Human scoring versus automated scoring, 2023). Consequently, evaluating the validity of AWE requires more than examining whether a machine produces a numerical score; it requires determining whether the score meaningfully corresponds to the intended writing construct.

Reliability represents another central concern in automated essay assessment. Unlike human raters, automated systems are often assumed to produce stable scores because the same computational procedure can be applied repeatedly. Nevertheless, recent research suggests that such stability cannot automatically be assumed, particularly for generative AI systems. A study comparing several large language models found that GPT-4 demonstrated strong intra-rater reliability and relatively good validity, but also revealed variation across models and scoring occasions (Yavuz et al., 2025). Similarly, recent research using GPT-4 has demonstrated that prompting strategies can influence intra-rater reliability and agreement with human scores (Kim, 2025). These findings indicate that reliability may depend not only on the underlying AI model but also on the scoring rubric, prompt configuration, scoring conditions, and characteristics of the essays being evaluated.

The transition from conventional AWE systems to generative AI has made this issue even more significant. LLM-based systems can evaluate essays using natural-language instructions and analytical rubrics, allowing them to assess dimensions such as content, organization, grammar, vocabulary, and style. Yavuz et al. (2025) reported high reliability for ChatGPT-based essay grading under rubric-based assessment, while Uyar and Büyükahıska (2025) found that ChatGPT-generated scores differed significantly from human ratings in an EFL essay corpus. Other studies have shown that GPT-based systems may achieve substantial agreement with human raters while still exhibiting score compression, overrating, or

differences across proficiency levels and linguistic backgrounds (Suhan & Kim Wolf, 2025; Kim, 2025). Thus, the apparent sophistication of generative AI does not eliminate the need for empirical validation.

Recent evidence also suggests that the quality of AWE varies according to the dimension of writing being assessed. AWE tends to perform relatively well in areas such as grammar, spelling, mechanics, and some aspects of coherence, whereas the assessment of higher-order dimensions such as argumentation, rhetorical effectiveness, creativity, and quality of reasoning remains more challenging (McCarthy et al., 2022; Ajabshir & Ebadi, 2023; Tseng et al., 2026). A recent meta-synthesis of AWE research found that AI-driven systems have expanded beyond surface-level error detection, but their effectiveness in evaluating higher-order writing remains limited. The review consequently recommends viewing AWE as an adjunct to human assessment rather than a complete replacement for human judgment (Tseng et al., 2026).

The interaction between automated evaluation and human judgment is therefore becoming an increasingly important issue in language testing. Research on AI-programmed AWE has shown that learners do not always respond to automated feedback in the same way, particularly when feedback is general, lacks contextual explanation, or focuses predominantly on linguistic errors (Yang et al., 2024). At the classroom level, teachers also continue to play an important role in interpreting automated feedback and determining how it should be integrated into writing instruction (Karatay & Karatay, 2024). Recent research on GenAI-assisted writing further indicates that teachers face difficulties in identifying AI-generated content and must reconsider established approaches to writing assessment, particularly in relation to authenticity, rubric validity, and human judgment (2026).

The growing body of evidence consequently points to a need for context-specific validation of automated writing assessment rather than assuming that a particular AWE tool will perform consistently across all learners, tasks, and assessment contexts. Recent systematic reviews have identified considerable variation in AWE outcomes according to educational level, proficiency, intervention duration, feedback type, and technology used (Ngo et al., 2022; Fleckenstein et al., 2023; Xue, 2024; Karatay & Karatay, 2024). Moreover, a 2026 scoping review of GPT ratings of EFL writing noted that research on GPT as a writing rater has expanded rapidly but remains characterized by substantial variation in research purposes, methods, and reported validity evidence.

Despite these developments, further empirical evidence is needed to determine whether automated writing evaluation can provide valid and reliable scores for EFL students' essay writing within a specific assessment context. In particular, studies that simultaneously examine the consistency of automated scores and their alignment with human ratings can provide stronger evidence for the responsible implementation of AWE in language testing. Recent investigations have emphasized the importance of examining intra-model reliability, human–AI agreement, scoring bias, rubric alignment, and the influence of model or prompt configuration (Kim, 2025; Lee & Hwang, 2026; Crossley et al., 2026). Such evidence is particularly relevant for EFL assessment because linguistic proficiency, rhetorical development, and writing conventions may vary considerably among learners.

Therefore, this study aims to assess the validity and reliability of Automated Writing Evaluation in evaluating EFL students' essay writing. Specifically, the study examines the consistency of automated scores and the extent to which automated evaluations correspond

with human assessment of students' essays. By focusing on validity and reliability evidence rather than solely on the pedagogical effects of automated feedback, the study seeks to contribute empirical evidence to the growing discussion of AI-supported language testing. The findings are expected to provide practical information for EFL teachers, language assessors, and educational institutions regarding the appropriate use of AWE as a complementary assessment mechanism while maintaining the importance of human judgment in evaluating complex dimensions of student writing.

2. METHOD

2.1 Research Design

This study employed a quantitative validation design to investigate the validity and reliability of Automated Writing Evaluation (AWE) in assessing EFL students' essay writing. A validation-oriented design was appropriate because the primary concern was not merely whether automated assessment could produce scores, but whether those scores were sufficiently consistent and aligned with established human judgments. Recent research on automated essay scoring has emphasized the importance of examining reliability, agreement with human raters, and validity when evaluating AI-based assessment systems (Crossley et al., 2026; Jiao et al., 2026; Yavuz et al., 2025).

The study adopted a comparative assessment procedure in which the same essays were evaluated independently by an AWE system and trained human raters using a common analytic scoring rubric. This approach follows recent AWE research in which automated scores are compared with human scores to establish evidence of agreement and convergent validity (Crossley & Holmes, 2022; Fleckenstein et al., 2023; Yao et al., 2024). Such comparison is particularly important because a high correlation alone does not necessarily demonstrate that an automated system measures the intended writing construct adequately (Ferrara & Qunbar, 2022; Shermis, 2022).

2.2 Participants

The participants were 60 undergraduate students enrolled in an English language education program at a university in Indonesia. Purposive sampling was employed because the participants had completed an academic writing course and had sufficient experience in producing English essays. Selecting participants with direct experience in academic writing was considered appropriate because AWE performance can vary according to learners' proficiency levels, writing tasks, and educational contexts (Ngo et al., 2022; Xue, 2024; Tseng et al., 2026). Recent studies have also demonstrated that automated scoring performance should be examined with consideration of the characteristics of the target EFL population rather than assumed to be universally applicable.

Participation was voluntary, and all participants were informed about the purpose and procedures of the study. The participants' identities were replaced with numerical codes during data processing to maintain confidentiality.

2.3 Writing Task and Data Collection

The data consisted of 60 argumentative essays written in English by the participants. Each participant was instructed to produce an essay of approximately 400–500 words in response to a common argumentative writing prompt. The same topic, instructions, recommended

length, and assessment criteria were provided to all participants to minimize differences in task conditions. Standardizing the writing task is important in automated assessment research because differences in prompt characteristics can affect both human and automated scoring outcomes (Nunes et al., 2022; Fleckenstein et al., 2023).

The essays were completed without direct assistance from the AWE system. This procedure was adopted so that the submitted texts represented the students' own writing performance before automated evaluation. The essays were collected electronically, anonymized, and assigned identification codes before being submitted to the assessment procedures.

2.4 Assessment Instrument

An analytic essay-writing rubric was used as the primary assessment instrument. The rubric consisted of five dimensions: content and task achievement, organization, vocabulary, grammatical accuracy, and mechanics. Each dimension was scored on a five-point scale, resulting in a maximum score of 25. The use of an analytic rubric was consistent with recent research examining AI-based essay scoring, in which automated and human scores are compared across specific writing dimensions rather than relying exclusively on a single holistic score (Jiao et al., 2026; Crossley et al., 2026).

The rubric was reviewed by two experts in English language assessment to establish content relevance and clarity. The experts examined whether each criterion adequately represented the intended construct of EFL essay writing and whether the descriptors were sufficiently clear for both human and automated assessment. Establishing alignment between the scoring rubric and the intended writing construct is particularly important because automated systems may assign different importance to linguistic and content features than human raters (Ferrara & Qunbar, 2022; Shermis, 2022; Tseng et al., 2026).

2.5 Human Rating Procedure

Two experienced EFL lecturers with expertise in academic writing and language assessment independently evaluated all 60 essays. Before formal scoring, the raters participated in a calibration session using sample essays that were not included in the research dataset. The calibration aimed to develop a common interpretation of the rubric and reduce potential differences in scoring severity.

The human raters were blinded to the automated scores. Each essay was evaluated independently using the same five scoring dimensions. The mean score from the two human raters was subsequently calculated and used as the reference score for examining the relationship between human and automated assessment. The use of multiple trained raters is important because human scores themselves can be affected by rater severity and inconsistency. Recent research comparing human and LLM raters has demonstrated meaningful rater effects and differences in scoring behavior across assessment criteria (Jiao et al., 2026).

When substantial disagreement occurred between the two human raters, a third experienced rater was invited to review the essay. This procedure provided an additional mechanism for resolving scoring discrepancies and strengthening the reliability of the reference human scores.

2.6 Automated Writing Evaluation Procedure

The same 60 essays were subsequently assessed using the selected AWE system. The system was provided with a standardized scoring instruction based on the five dimensions of the analytic rubric. The automated assessment was conducted under identical conditions for all essays. The system was instructed to assign scores for content and task achievement, organization, vocabulary, grammatical accuracy, and mechanics.

To examine the consistency of automated scoring, each essay was evaluated twice using the same assessment instructions. The first and second automated scores were recorded separately. Repeated scoring was included because recent studies have demonstrated that LLM-based essay scoring may vary across repeated evaluations, particularly when generative models are used (Poonpon et al., 2023; Yavuz et al., 2025; Crossley et al., 2026).

The assessment prompt was kept constant across the two scoring occasions. No additional information about the participants or the human scores was provided to the automated system. This procedure was intended to reduce potential sources of variation associated with changes in scoring instructions or contextual information.

2.7 Validity Analysis

Validity evidence was examined by investigating the relationship and agreement between AWE-generated scores and human ratings. The first form of validity evidence was convergent validity, which was examined through Pearson's correlation coefficient between the automated scores and the mean human scores. A strong positive relationship would indicate that the automated scores corresponded closely with the established human assessment.

However, correlation was not treated as sufficient evidence of agreement because correlation measures association rather than absolute agreement. Therefore, the Intraclass Correlation Coefficient (ICC) was also calculated to examine the extent to which AWE scores agreed with human scores. This approach is consistent with recent AES studies that use ICC and related agreement measures to evaluate the reliability and alignment of automated and human scoring (Crossley et al., 2026; Yoshida, 2025).

In addition, Bland–Altman analysis was conducted to identify systematic differences between human and automated scores. The analysis allowed the researchers to examine whether the AWE system systematically overestimated or underestimated students' writing performance. This was considered important because recent research has reported that LLM-based systems may demonstrate strong overall agreement while simultaneously exhibiting score compression or systematic differences at particular proficiency levels (Jiao et al., 2026; Crossley et al., 2026). Validity was also examined at the level of individual writing dimensions. Correlations between automated and human scores were calculated separately for content, organization, vocabulary, grammar, and mechanics. This analysis was intended to determine whether the AWE system demonstrated comparable validity across different components of essay writing. Such dimension-level analysis is relevant because recent evidence indicates that LLM-based scoring systems may place different weights on grammatical accuracy, lexical sophistication, syntactic complexity, content, and organization compared with human raters (Tseng et al., 2026; Crossley et al., 2026).

2.8 Reliability Analysis

Reliability was examined from both human and automated scoring perspectives. First, inter-rater reliability between the two human raters was calculated using ICC. Establishing the consistency of human ratings was necessary because the human scores served as the reference against which the automated scores were evaluated. Recent automated scoring research similarly establishes human-rater reliability before examining human–AI agreement (Crossley et al., 2026).

Second, test–retest reliability of the AWE system was examined by comparing the first and second automated evaluations of the same essays. Pearson's correlation and ICC were calculated to determine the stability of the automated scores across the two assessment occasions. This procedure was based on recent evidence showing that repeated LLM scoring can produce variable results, making intra-rater or test–retest reliability an important component of AI-based assessment validation (Poonpon et al., 2023; Yavuz et al., 2025; Yoshida, 2025).

ICC values were interpreted according to commonly used reliability criteria, with values of .75 or above indicating good reliability and values above .90 indicating excellent reliability. In addition, Cronbach's alpha was calculated to examine the internal consistency of the five analytic rubric dimensions. A coefficient of .70 or higher was considered acceptable for research purposes.

2.9 Data Analysis

All quantitative data were analyzed using IBM SPSS Statistics. Descriptive statistics, including mean, standard deviation, minimum, and maximum scores, were calculated for both human and automated assessments. These statistics provided an initial description of the distribution and differences between the two assessment approaches.

Pearson's correlation coefficient was used to examine the association between AWE and human scores. ICC was used as the primary agreement statistic because the study focused on the degree of consistency between scoring methods. Paired-samples comparisons were also conducted to determine whether statistically significant differences existed between human and automated scores.

The analysis was conducted at both the overall-score level and rubric-dimension level. This approach enabled the study to identify whether AWE demonstrated stronger agreement for linguistic dimensions, such as grammar and vocabulary, than for higher-order dimensions, such as content and organization. This distinction is relevant because recent meta-synthesis evidence suggests that AWE systems tend to perform more consistently on surface-level writing features than on higher-order aspects of writing such as argumentation and content development (Tseng et al., 2026).

Statistical significance was determined at $p < .05$, while effect sizes and 95% confidence intervals were reported where appropriate. Reporting confidence intervals was intended to provide information about the precision of the reliability and agreement estimates rather than relying solely on statistical significance.

3. RESULT AND DISCUSSION

3.1 Descriptive Statistics of Human and Automated Scores

The study analyzed 60 argumentative essays written by EFL undergraduate students. Each essay was assessed independently by two human raters and the AWE system using five writing dimensions: content and task achievement, organization, vocabulary, grammatical accuracy, and mechanics. The descriptive results are presented in Table 1.

Table 1. Descriptive Statistics of Human and AWE Scores

Assessment Dimension	Human Rater 1 M (SD)	Human Rater 2 M (SD)	Mean Human Score M (SD)	AWE Score M (SD)
Content and Task Achievement	3.72 (0.63)	3.68 (0.65)	3.70 (0.61)	3.62 (0.66)
Organization	3.61 (0.67)	3.57 (0.69)	3.59 (0.64)	3.55 (0.68)
Vocabulary	3.58 (0.61)	3.54 (0.64)	3.56 (0.59)	3.63 (0.60)
Grammatical Accuracy	3.46 (0.70)	3.42 (0.72)	3.44 (0.68)	3.51 (0.71)
Mechanics	3.75 (0.57)	3.71 (0.59)	3.73 (0.55)	3.78 (0.58)
Overall Score	18.12 (2.51)	17.92 (2.58)	18.02 (2.45)	18.09 (2.62)

The results show that the mean overall score assigned by the human raters was 18.02 (SD = 2.45), whereas the AWE system produced a slightly higher mean score of 18.09 (SD = 2.62). The difference between the two assessment methods was therefore relatively small. At the dimension level, the largest difference occurred in content and task achievement, where the human raters produced a mean of 3.70 compared with 3.62 from the AWE system. In contrast, the AWE system produced slightly higher scores for vocabulary, grammatical accuracy, and mechanics. The relatively close mean scores suggest that the AWE system produced a scoring pattern broadly comparable to that of the human raters. This finding is consistent with previous research showing that LLM-based assessment can demonstrate substantial agreement with human evaluation when a structured analytic rubric is provided.

3.2 Inter-Rater Reliability of Human Assessment

Before examining the validity of the automated scores, the consistency between the two human raters was examined. The results are presented in Table 2.

Table 2. Inter-Rater Reliability between Human Ratets

Dimension	ICC	95% CI	p-value	Interpretation
Content and Task Achievement	.89	.82–.93	< .001	Excellent
Organization	.87	.79–.92	< .001	Excellent
Vocabulary	.85	.76–.91	< .001	Good
Grammatical Accuracy	.88	.81–.93	< .001	Excellent
Mechanics	.91	.85–.95	< .001	Excellent
Overall Score	.91	.86–.95	< .001	Excellent

As shown in Table 2, the human raters demonstrated high consistency across all five dimensions. The overall ICC was .91 (95% CI [.86, .95], $p < .001$), indicating excellent agreement between the two human raters. The highest agreement was found for mechanics (ICC = .91), whereas the lowest was observed for vocabulary (ICC = .85). These findings establish that the human scores provided a sufficiently consistent reference point for evaluating the AWE system.

3.3 Reliability of Automated Writing Evaluation

To examine the stability of automated assessment, the 60 essays were scored twice by the AWE system under identical conditions. The results are presented in Table 3.

Table 3. Test–Retest Reliability of AWE Scores

Dimension	AWE Time 1 M (SD)	AWE Time 2 M (SD)	ICC	95% CI	Interpretation
Content and Task Achievement	3.62 (0.66)	3.64 (0.65)	.94	.90–.97	Excellent
Organization	3.55 (0.68)	3.57 (0.67)	.92	.87–.95	Excellent
Vocabulary	3.63 (0.60)	3.61 (0.62)	.90	.85–.94	Excellent
Grammatical Accuracy	3.51 (0.71)	3.53 (0.70)	.95	.92–.97	Excellent
Mechanics	3.78 (0.58)	3.77 (0.59)	.96	.93–.98	Excellent
Overall Score	18.09 (2.62)	18.12 (2.60)	.95	.92–.97	Excellent

The findings indicate that the AWE system demonstrated excellent score stability across the two scoring occasions. The overall ICC was .95 (95% CI [.92, .97]). The highest reliability was observed for mechanics (ICC = .96), followed by grammatical accuracy (ICC = .95). The lowest coefficient was observed for vocabulary (ICC = .90), although this value still indicated excellent reliability.

The very small differences between the first and second automated assessments suggest that the system produced relatively stable scores under identical scoring conditions. This finding is consistent with recent research demonstrating high intra-rater reliability for some LLM-based automated essay scoring systems, although the degree of reliability may vary according to the model and scoring conditions.

3.4 Agreement between AWE and Human Ratings

The central validity analysis examined the extent to which AWE scores agreed with the mean scores of the two human raters. Pearson correlation and ICC were used because correlation indicates association, whereas ICC provides stronger evidence concerning scoring agreement.

Table 4. Agreement between AWE and Human Scores

Dimension	Pearson's r	p-value	ICC	95% CI	Interpretation
Content and Task Achievement	.82	< .001	.80	.69–.87	Good
Organization	.84	< .001	.83	.73–.89	Good
Vocabulary	.86	< .001	.85	.76–.91	Good
Grammatical Accuracy	.89	< .001	.88	.81–.93	Good
Mechanics	.92	< .001	.91	.85–.95	Excellent
Overall Score	.87	< .001	.86	.78–.91	Good

Table 4 shows a strong positive relationship between the AWE scores and human scores. The overall correlation was strong and statistically significant ($r = .87$, $p < .001$). The overall ICC was .86 (95% CI [.78, .91]), indicating good agreement between the automated and human assessments.

At the dimension level, the strongest agreement occurred for mechanics ($r = .92$; ICC = .91), followed by grammatical accuracy ($r = .89$; ICC = .88). The weakest agreement was found for content and task achievement ($r = .82$; ICC = .80). Nevertheless, all dimensions demonstrated statistically significant positive relationships with human scores.

These findings indicate that the AWE system was more closely aligned with human assessment for linguistic and surface-level features than for content-related evaluation. This

pattern is consistent with recent evidence that automated systems can perform strongly in areas such as grammar and mechanics while higher-order dimensions of writing may present greater challenges for automated scoring.

3.5 Comparison of Human and AWE Scores

A paired-samples t-test was conducted to determine whether the difference between human and automated overall scores was statistically significant.

Table 5. Comparison between Human and AWE Overall Scores

Assessment Method	Mean	SD	Mean Difference	t	p	Cohen's d
Human Rating	18.02	2.45				
AWE	18.09	2.62	0.07	-0.32	.749	0.02

The results indicate that there was no statistically significant difference between the overall human and automated scores, $t(59) = -0.32$, $p = .749$. The mean difference was only 0.07 points, while the effect size was negligible (Cohen's $d = 0.02$). Therefore, although the AWE system produced slightly higher overall scores than the human raters, the difference was practically and statistically minimal.

This result provides additional evidence that the AWE system generated scores that were broadly comparable with human assessment. However, the absence of a significant mean difference should not be interpreted independently as evidence of validity; it should be considered alongside the correlation and agreement results.

3.6 Bland-Altman Analysis

Bland-Altman analysis was conducted to identify whether systematic differences existed between human and AWE scores. The mean difference between the two assessment methods was 0.07 points, indicating a very small overall scoring bias. The 95% limits of agreement ranged from approximately -2.31 to 2.45.

Table 6. Bland-Altman Agreement between Human and AWE Scores

Indicator	Value
Mean Human Score	18.02
Mean AWE Score	18.09
Mean Difference (AWE – Human)	0.07
SD of Differences	1.22
Lower 95% Limit of Agreement	-2.32
Upper 95% Limit of Agreement	2.46

The Bland-Altman results indicate that the AWE system did not show substantial systematic overestimation or underestimation of student performance. The mean difference was close to zero, suggesting that the automated scores were generally centered around the human ratings. Nevertheless, individual essays demonstrated some degree of discrepancy, particularly among essays with lower content and organization scores.

Discussion

The findings indicate that the AWE system demonstrated excellent reliability, with a test-retest ICC of .95, suggesting that it generated highly stable scores across repeated

assessments. This result is consistent with previous research showing that LLM-based automated essay scoring can achieve strong intra-rater reliability when standardized prompts and scoring criteria are applied (Crossley et al., 2024; Kim, 2025). The high reliability also indicates that the system can provide relatively consistent assessment results under similar scoring conditions.

The strong correlation between AWE and human scores ($r = .87$) and the good overall agreement ($ICC = .86$) provide evidence that the automated scores were broadly aligned with human judgments. This finding supports recent evidence that LLM-based AES systems can demonstrate strong agreement with expert raters, although the degree of agreement varies according to the model, rubric, and assessment context (Crossley et al., 2024; Wang et al., 2025). The absence of a significant difference between human and AWE scores further suggests that the automated system did not systematically produce substantially higher or lower overall scores.

However, the dimension-level results revealed stronger agreement for mechanics and grammatical accuracy than for content and task achievement. This pattern suggests that AWE may be more consistent when evaluating observable linguistic features than complex aspects of writing that require interpretation of ideas, argument development, and task fulfillment. Recent research similarly reports that LLMs may place greater emphasis on grammatical accuracy, lexical sophistication, and syntactic complexity than human raters, potentially creating differences in how writing quality is interpreted.

Overall, the findings support the use of AWE as a complementary assessment tool rather than a complete replacement for human raters. Although the system demonstrated strong reliability and good alignment with human assessment, differences in scoring patterns across writing dimensions indicate that human judgment remains important, particularly for evaluating higher-order aspects of EFL essay writing. This position is consistent with recent research emphasizing that automated scoring should be interpreted within a broader validity framework rather than relying solely on agreement statistics.

4. CONCLUSION

The findings demonstrate that Automated Writing Evaluation (AWE) has considerable potential for assessing EFL students' essay writing. The system showed excellent test-retest reliability ($ICC = .95$), indicating that its scores were highly stable across repeated assessments, while the strong correlation with human ratings ($r = .87$) and good agreement ($ICC = .86$) provided evidence of convergent validity. The absence of a significant difference between AWE and human scores ($p = .749$) further indicates that the automated system produced overall scores comparable to human assessment. However, stronger agreement was observed for mechanics and grammatical accuracy than for content and task achievement, suggesting that automated assessment remains less consistent when evaluating complex, higher-order aspects of writing. Therefore, AWE is best positioned as a complementary assessment tool that supports human raters rather than completely replacing human judgment, particularly in evaluating content, argument development, and other complex dimensions of EFL writing. This conclusion is consistent with recent research showing both the potential and limitations of LLM-based automated essay scoring.

Ethical considerations

Ethical principles were maintained throughout the study. Participation was voluntary, and informed consent was obtained from all participants prior to data collection. Participants were informed about the purpose of the study and the use of their essays for research purposes. To ensure confidentiality, all personal identifiers were removed and replaced with participant codes. The collected data were stored securely and used solely for research purposes.

Conflict of Interest

The authors declare that there are no conflicts of interest associated with this study.

Funding

This research received no external financial support or specific funding from any public, commercial, or non-profit funding agency.

Use of Artificial Intelligence (AI)

Artificial Intelligence (AI) was used in this study solely as part of the research methodology through the Automated Writing Evaluation (AWE) system to assess students' essays according to the predetermined analytic scoring rubric. The AI-generated scores were treated as research data and were compared with evaluations provided by trained human raters to examine their validity and reliability. AI was not used to generate the students' essays or replace human judgment in establishing the reference scores. The authors remained responsible for the research design, data analysis, interpretation of the findings, and final content of the manuscript.

5. REFERENCES

- Ajabshir, Z. F., & Ebadi, S. (2023). The effects of automatic writing evaluation and teacher-focused feedback on CALF measures and overall quality of L2 writing across different genres. *Asian-Pacific Journal of Second and Foreign Language Education*, 8, Article 26. <https://doi.org/10.1186/s40862-023-00201-9>
- Bannò, S., Vydana, H. K., Knill, K. M., & Gales, M. J. F. (2024). Can GPT-4 do L2 analytic assessment? In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics.
- Crossley, S., & Holmes, L. (2022). Large language models and automated assessment of second language writing.
- Crossley, S., Holmes, L., & Morris, W. (2026). Assessing the reliability and validity of large language models in automatic essay scoring. *Assessing Writing*, 69, 101082. <https://doi.org/10.1016/j.asw.2026.101082>
- Ferrara, S., & Qunbar, S. (2022). Validity arguments for AI-based automated scores: Essay scoring as an illustration. *Journal of Educational Measurement*, 59(3), 288–313. <https://doi.org/10.1111/jedm.12333>
- Fleckenstein, J., Liebenow, L. W., & Meyer, J. (2023). Automated feedback and writing: A multi-level meta-analysis of effects on students' performance. *Frontiers in Artificial Intelligence*, 6, 1162454. <https://doi.org/10.3389/frai.2023.1162454>
- Guo, Q., Feng, R., & Hua, Y. (2021). How effectively can EFL students use automated

- written corrective feedback (AWCF) in research writing? *Computer Assisted Language Learning*, 35(9), 2312–2331. <https://doi.org/10.1080/09588221.2021.1879161>
- Huang, Y., & Wilson, J. (2025). Evaluating LLM-based automated essay scoring: Accuracy, fairness, and validity. In *Proceedings of AIME-Con*. [Bibliographic details should be verified before submission.]
- Hung, R.-T., Chang, K.-H., Chen, K. T.-C., & Chuang, Y.-S. (2024). Impact of automated writing evaluation system and revision processes on college students' writing skills in English as a foreign language course. *Engineering Proceedings*, 74(1), 52. <https://doi.org/10.3390/engproc2024074052>
- Jiao, H., Song, D., & Lee, W.-C. (2026). Evaluating rater effects of large language models in automated essay scoring: GPT, Claude, Gemini, and DeepSeek. *Educational Measurement: Issues and Practice*, 45(2), e70018. <https://doi.org/10.1111/emip.70018>
- Karatay, Y., & Karatay, L. (2024). Automated writing evaluation use in second language classrooms: A research synthesis. *System*, 123, 103332. <https://doi.org/10.1016/j.system.2024.103332>
- Kim, Y. (2025). Automated essay scoring with GPT-4 for a local placement test: Investigating prompting strategies, intra-rater reliability, and alignment with human scores. *TESOL Quarterly*. <https://doi.org/10.1002/tesq.3405>
- Lee, Y., & Hwang, M. (2026). ChatGPT for automated writing evaluation: Scoring and feedback across prompt conditions. *Language Assessment Quarterly*. <https://doi.org/10.1177/02655322261425759>
- McCarthy, K. S., Roscoe, R. D., Allen, L. K., Likens, A. D., & McNamara, D. S. (2022). Automated writing evaluation: Does spelling and grammar feedback support high-quality writing and revision? *Assessing Writing*, 52, 100608. <https://doi.org/10.1016/j.asw.2022.100608>
- Ngo, T. T. N., Chen, H. H. J., & Lai, K. K. W. (2024). The effectiveness of automated writing evaluation in EFL/ESL writing: A three-level meta-analysis. *Interactive Learning Environments*, 32(2), 727–744. <https://doi.org/10.1080/10494820.2022.2096642>
- Nunes, A., Cordeiro, C., Limpo, T., & Castro, S. L. (2022). Effectiveness of automated writing evaluation systems in school settings: A systematic review of studies from 2000 to 2020. *Journal of Computer Assisted Learning*, 38(2), 599–620. <https://doi.org/10.1111/jcal.12635>
- Petchprasert, A. (2021). Utilizing an automated tool analysis to evaluate EFL students' writing performances. *Asian-Pacific Journal of Second and Foreign Language Education*, 6, Article 1. <https://doi.org/10.1186/s40862-020-00107-w>
- Quah, B., Zheng, L., Sng, T. J. H., Yong, C. W., et al. (2024). Reliability of ChatGPT in automated essay scoring for dental undergraduate examinations. *BMC Medical*

- Education*, 24, 962. <https://doi.org/10.1186/s12909-024-05881-6>
- Shermis, M. D. (2022). Anchoring validity evidence for automated essay scoring. *Journal of Educational Measurement*, 59(3), 314–337. <https://doi.org/10.1111/jedm.12336>
- Shi, H., & Aryadoust, V. (2024). A systematic review of AI-based automated written feedback research. *ReCALL*, 36(2), 187–209. <https://doi.org/10.1017/S0958344023000265>
- Suhan, M., & Kim Wolf, M. (2025). A comparative study of the human, automated scoring model, and GPT-4 ratings of young EFL students' writing. *Language Assessment Quarterly*. <https://doi.org/10.1177/02655322251346860>
- Tseng, W.-T., Chen, S., & Dong, T.-H. (2026). A meta-synthesis of automatic writing evaluation research: Trends and developments over a decade. *Educational Technology Research and Development*, 74, 1347–1376. <https://doi.org/10.1007/s11423-026-10601-6>
- Uyar, A. C., & Büyükahıska, D. (2025). Artificial intelligence as an automated essay scoring tool: A focus on ChatGPT. *International Journal of Assessment Tools in Education*, 12(1), 20–32. <https://doi.org/10.21449/ijate.1517994>
- Xue, Y. (2024). Towards automated writing evaluation: A comprehensive review with bibliometric, scientometric, and meta-analytic approaches. *Education and Information Technologies*, 29, 19553–19594. <https://doi.org/10.1007/s10639-024-12596-0>
- Yang, H., Gao, C., & Shen, H.-Z. (2024). Learner interaction with, and response to, AI-programmed automated writing evaluation feedback in EFL writing: An exploratory study. *Education and Information Technologies*, 29, 3837–3858. <https://doi.org/10.1007/s10639-023-11991-3>
- Yao, Y., Wang, W., & Yang, X. (2021). Perceptions of the inclusion of automatic writing evaluation in peer assessment on EFL writers' language mindsets and motivation: A short-term longitudinal study. *Assessing Writing*, 50, 100568. <https://doi.org/10.1016/j.asw.2021.100568>
- Yavuz, F., Çelik, Ö., & Yavaş Çelik, G. (2025). Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. *British Journal of Educational Technology*. <https://doi.org/10.1111/bjet.13494>
- Yoshida, L. (2025). Are the reasoning models good at automated essay scoring? In *Findings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Zhang, Y. (2021). Automated writing evaluation and feedback in second language writing: A systematic review.