Journal of Posthumanism
2025
Volume: 5, No: 6, pp. 1533–1549
ISSN: 2634-3576 (Print) | ISSN 2634-3584 (Online)
posthumanism.co.uk

DOI: https://doi.org/10.63332/joph.v5i6.2224

# Utilizing Artificial Intelligence to Assess ESL Students' Narratives: A Comparative Analysis

Asmaa Alshehri[1], Salha Hassan Qahl[2], Najlaa Alharbi[3]

***Abstract***

*This study investigates the effectiveness, reliability, and potential biases of AI-based assessment tools in evaluating narrative essays written by undergraduate ESL students at a Saudi university. A total of 30 essays were assessed using a detailed rubric covering five writing components: ideas and content, organization, vocabulary, voice and style, and mechanics and formatting. The essays were graded by human evaluators and five AI tools—ChatGPT, Gemini, Claude, Justdone, and Chatsonic. A quantitative comparative research design was employed, and statistical analyses, including one-way ANOVA and correlation tests, were conducted to examine grading consistency and divergence. Results revealed that AI tools aligned more closely with human graders on objective criteria like mechanics and formatting, but showed significant discrepancies in subjective aspects such as voice and style. The study highlights the potential of AI to support human grading but underscores the importance of human oversight to ensure fairness and contextual sensitivity in ESL writing assessment.*

***Keywords:*** *Artificial Intelligence (AI), Assessment, ESL Narrative Writing, Comparative Analysis*

## Introduction

The integration of Artificial Intelligence (AI) into educational assessment has become a burgeoning area of interest, with a growing body of research exploring its impact on student outcomes, educator practices, and ethical implications. Sánchez-Prieto et al. (2020) provided a comprehensive categorization of AI's role in educational assessment. However, they identified a critical gap in the literature: the pedagogical implications of AI in student assessment remain underexamined. This gap underscores the need for further research, particularly particularly in English as a second language (ESL) literary writing contexts, such as in Saudi universities. The increasing prevalence of AI tools in education necessitates a deeper understanding of their effectiveness across different types of assessment tasks to properly integrate these technologies into meaningful pedagogical frameworks.

Several studies have begun exploring different aspects of AI's impact on student assessment. For example, Nazari et al. (2021) conducted an empirical study on AI's role in academic writing, particularly for non-native graduate students. They found that AI tools such as Grammarly significantly enhanced behavioral, emotional, and cognitive engagement as it emphasizes AI's positive role in providing formative feedback and assessment. This aligns closely with the current

[1] Department of English Language and Literature University of Bisha 255, Al Nakhil Bisha, Asir 67714 Saudi Arabia, Phone: +966 552 354 158, Email: afadel@ub.edu.sa
[2] Medical Sciences Graduate Programs (Foothills Campus) University of Calgary TRW 2D08, 3280 Hospital Drive NW Calgary, AB T2N 4Z6 Canada Phone: +1 403 210 8764, Email: Salha.Salha@ucalgary.ca
[3] Department of English Language and Literature Qassim University 6688, Buraydah Al-Qassim 51452 Saudi Arabia, Phone: +966 505 321 649, Email: nr.alharbi@qu.edu.sa.

research focus on evaluating the accuracy and efficacy of AI tools in ESL contexts. The formative feedback contributes to a more dynamic and effective learning environment. Nazari et al.'s (2021) findings are particularly relevant to the current discussion, as they emphasize AI's potential to improve writing outcomes and facilitate a more nuanced approach to student learning. This aligns closely with the focus of this research on ESL contexts in Saudi universities. However, it is essential to assess not only the short-term gains from using AI tools but also their accuracy and reliability in specific assessment contexts such as ESL literary writing.

Building on this, Rahman et al. (2023) offered a more critical view, focusing on the specific capabilities and limitations of AI in language assessment. While Grammarly was shown to enhance language mechanics—grammar, syntax, and spelling—Rahman et al. (2023) highlighted its limitations in evaluating deeper elements of writing, such as content and organizational skills. This finding reinforces the importance of human oversight in assessing higher-order tasks. Its implications are crucial for this study, as they underline the limitations of relying solely on AI tools for complex assessments and stress the need for combining human and AI-based grading approaches.

In contrast, ethical concerns about AI in educational contexts have been extensively discussed. Naidu and Sevnarayan (2023) expressed caution regarding the biases that AI tools may perpetuate, as well as the potential for diminished human interaction in learning environments. They warned that overreliance on AI in grading, especially when using tools such as ChatGPT, could marginalize the role of human assessors, leading to issues of fairness and transparency in student evaluations. This perspective invites a broader critical reflection on AI's role, suggesting that while AI can provide significant benefits in efficiency and scalability, its use in high-stakes assessment should be approached with care. The current study's exploration of AI-based grading systems versus human evaluation directly engages with these ethical concerns, seeking to balance the efficiencies of AI with the need for fairness and human oversight.

The balance between AI efficiency and human oversight is also emphasized by Mao et al. (2024), who advocated for a human-centric approach to AI integration in education. Their research on the ethical challenges posed by AI, as well as the importance of developing AI literacy among educators, is particularly relevant to this present research. The study underscores the need for educators to be equipped not just with AI tools but also with the knowledge to use them critically and responsibly. This aligns with the goals of this present research, goals which aim to critically assess the fairness, accuracy, and reliability of AI tools in ESL assessment contexts.

Complementing these broader perspectives, Ali (2023) explored the impact of training in AI on teachers' perceptions regarding AI in ESL classrooms. Ali (2023) found that structured AI training positively shifted teachers' attitudes. The results by Ali (2023) also showed that if integrated appropriately and with adequate training, AI tools hold great promise for improving teaching and assessment practices—a fact relevant to this study, which aims to develop a practical application of AI into the assessment of ESL literary writing. This speaks even more to the correlation between teacher training and institutional support and how those are necessary in ways that are effective and intentional in using AI tools.

Further nuance is added by studies from Anuyahong, Rattanapong, and Patcha (2023) and Mahapatra (2024), which explore both the potential benefits and limitations of AI in education. Mahapatra's (2024) study on ChatGPT's role in improving ESL students' writing skills demonstrated that AI-driven feedback could enhance writing proficiency but also pointed out the mechanistic nature of the feedback and its reliance on the user's existing proficiency. These

findings resonate with the current study's emphasis on comparing AI-generated feedback with human assessments, particularly in the context of ESL literary writing where more interpretative and creative aspects of writing need to be evaluated—areas where AI may struggle.

Finally, Liao, Xiao, and Hu (2023) provided an in-depth examination of ChatGPT's performance against established writing standards, noting that while its feedback was generally accurate, it often lacked the nuanced understanding that human evaluators offer. This aligns with the broader narrative of the current study, which investigates the comparative reliability and accuracy of AI-based automated essay scoring (AES) tools versus human graders. The acknowledgment that AI cannot fully grasp subtle aspects of human writing—such as tone, creativity, and rhetorical complexity—is crucial for understanding the limitations of AI in literary assessment.

In summary, the literature highlights both the potential and limitations of AI in educational settings. While AI tools such as Grammarly and ChatGPT have been shown to improve mechanical aspects of writing and foster student engagement, there remain significant concerns about their ability to accurately assess higher-order cognitive tasks, including content development, coherence, and creativity. Former studies such as those by Mao et al. (2024) and Rahman et al. (2023) suggest that human oversight is crucial in ensuring that AI tools complement rather than replace the nuanced judgment required in complex assessments. Ethical concerns regarding fairness, bias, and the diminished role of human interaction in learning environments also underscore the importance of a balanced approach to AI integration. This necessitates ongoing research into hybrid assessment models that combine AI efficiency with human expertise. As AI becomes increasingly integrated into educational systems, future studies, particularly in ESL contexts, must focus on refining AI-based assessments to enhance their reliability, fairness, and pedagogical value.

### Research Questions

The research questions guiding the current research are:

*1. To what extent do human grader and AI evaluations of students' essays converge or diverge, considering both overall scores and specific rubric criteria?*

*2. How do the inter-criterion correlation matrices generated by AI tools compare with those derived from human grader assessments?*

*3.*

### Research Hypotheses

The hypotheses addressed in this study are:

*Hypothesis 1: In evaluating ESL student essays, there is no significant difference between the overall scores given by human graders and AI tools.*

*Hypothesis 2: The correlation strengths in AI evaluations and those in human assessments show minimal differences across the various criteria of the rubric.*

## Methodology

### Research Design

This study adapted a quantitative comparative research design to analyze and compare the overall scores assigned to undergraduate students' narrative essays by human graders and various AI

tools. As Creswell and Creswell (2018) note, a quantitative comparative research design is effective for systematically analyzing relationships between variables while helping to identify patterns of agreement and disparity among different groups, conditions, and assessment methods. Moreover, as Babbie (2020) points out, this approach aligns with the principles of quantitative research by emphasizing objectivity, measurement, and statistical analysis. A quantitative comparative research design offers several advantages, such as providing objective, measurable comparisons that reduce potential bias and subjectivity during the evaluation process. Additionally, using a standardized rubric ensures consistency and comparability across different graders, whether human or AI. Statistical analysis helps to identify significant differences, further enhancing the reliability of the results.

## Research Context and Sampling

The writing samples for this study were drawn from undergraduate students majoring in English Language and Translation at a public university in Saudi Arabia. At the time of data collection, the students were enrolled in a three-credit academic writing course, meeting twice weekly for 90-minute sessions during the fall semester of 2023. As part of their coursework, the students composed essays in various genres, including argumentative, descriptive, and narrative essays. Toward the end of the semester, after submitting all their essays, the students were briefed on the research study's purpose and encouraged to share their narrative essays. Participation was entirely voluntary and did not affect course grades. Students were assured of the confidentiality of their identities and were given the right to withdraw their essays from the study at any time. Of the 42 students enrolled in the course, 35 agreed to share their essays and sign informed consent forms. However, five essays were excluded because the writers did not sign the consent forms, resulting in a final dataset of 30 essays. For analysis, each essay was anonymized and labeled 'NARRT' (narrative), with numbers assigned from 01 to 30 (e.g., NARRT_01 refers to the first student's essay).

## Instruments

To address the research questions, this study utilized the following instruments: the narrative rubric and AI-powered tools.

## The Narrative Rubric

A narrative rubric was developed to assess essays by both human graders and AI tools. This rubric included five main categories: 'Ideas and Content,' 'Organization,' 'Vocabulary,' 'Voice and Style,' and 'Mechanics and Formatting.' Each category was rated on a four-point scale (see Appendix A). To establish the rubric's content validity, three experienced professors specializing in ESL assessment revised each criterion so that it would accurately capture the key components of narrative essay writing. According to the experts' comments, the final version of the rubric was refined.

## Preliminary Testing

To assess the reliability of the rubric, six essays were randomly selected from the narrative corpus and independently assessed. The inter-rater reliability for each category of the rubric was measured using Cohen's kappa coefficient. The results showed that the researchers' rating of the first rubric criterion, 'Ideas and Content,' yielded a $\kappa$ of 0.96. On the second criterion, 'Organization,' the scoring generated a $\kappa$ of 0.93. For the third criterion, 'Vocabulary,' the rating yielded a $\kappa$ of 0.90, and for the fourth category, 'Voice and Style,' the scoring produced a $\kappa$ of

0.91. The scoring produced a κ of 0.89 for the final criterion of the rubric, 'Mechanics and Formatting.' As Landis and Koch (1977) suggested, these κ statistics indicate a substantial level of agreement between raters. Therefore, the narrative rubric is reliable for grading the students' narrative essays.

**AI-Powered Tools**

The present study utilized five AI-powered tools that use complex language models based on natural learning processing (NLP) and machine learning algorithms: ChatGPT, Gemini, Claude, Justdone, and Chatsonic. These particular AI-powered tools were chosen for this study not only because they are widely recognized and accessible but also, as Holmes et al. (2019) affirmed, because they are effective in performing different NLP tasks. Moreover, Sun (2023) pointed out that these tools are useful in educational contexts for reliable assessment.

A structured training process was conducted to establish the validity and reliability of the scoring of AI-powered tools for narrative essays. The researchers developed training prompts to instruct each of the AI-powered tools in comprehending the criteria of the narrative rubric and the scoring process. To ensure the reliability of the AI-powered tools' scoring, the six narrative samples, which were used to construct the reliability of the narrative rubric, were inserted into each of the AI-powered tools. Cohen's kappa coefficient was used to measure the inter-rater reliability between the AI-powered tools.

The results showed that for the first rubric criterion, 'Ideas and Content,' the AI tools' scoring yielded a κ of 0.90. The scoring generated a κ of 0.93 for the second criterion, 'Organization,' and a κ of 0.94 for the third criterion, 'Vocabulary,' of the rubric. On the fourth criterion, 'Voice and Style,' the scoring generated a κ of 0.90, and for the final criterion of the rubric, 'Mechanics and Formatting,' the scoring produced a κ of 0.89. These κ results suggest a considerable level of agreement between AI-powered tools' ratings (Landis & Koch, 1977).

Furthermore, Cohen's kappa coefficient was used to compare the scores assigned by the researchers and the scores assigned by AI-powered tools. The results showed that both the first, 'Ideas and Content,' and second, 'Organization,' criteria of the rubric indicated that a κ value between the researchers and AI-tools' scores was 0.93. For the third criterion, 'Vocabulary,' the κ was 0.92. For the last two criteria of the rubric, 'Voice and Style' and 'Mechanics and Formatting,' the κ value was 0.90. In fact, these results indicate that the scores generated by the researchers were consistent with the scores produced by AI-powered tools across all the criteria of the narrative rubric.

**Data Analysis Procedures**

The data analysis was conducted in four stages.

**First Phase**

The initial phase was the preparation stage for the analysis of the collected data. All of the 30 narrative essays were transformed into a Word document and then uploaded to a protected Google folder, which was accessed only by researchers. For privacy and confidentiality, the information leading to students' identities was removed, and each essay was labeled with a particular identification, as mentioned previously.

**Second Phase**

The second phase is the scoring process of the students' narrative essays by five raters, including the researchers and two graduate students interested in L2 writing. After being trained on how to use the narrative rubric, all the raters independently scored the 30 narrative essays. After that, the scores were inserted into a Google Word document that was developed to track the scores of each category of the rubric.

**Third Phase**

The third phase involved the scoring process of the narrative essays with AI-powered tools. Each of the five raters monitored the scoring process by focusing on a specific AI-powered tool while scoring the 30 narrative essays. The researchers utilized the developed prompts to guide the AI-powered tools through the grading process. Then, the scores generated by the AI-powered tools were entered into a designated table in a Google Word document for another round of analysis. In this phase, the researchers decided to exclude one essay (NARRT_13) from the dataset (N = 30). Making this critical decision was related to the fact that one of the AI-powered tools rejected assessing that essay because it was about a sensitive topic related to a student's life experience of sexual harassment. While the researchers acknowledge the potential impact of excluding this essay, they prioritized ethical considerations of the study over keeping data content that might be problematic.

**Fourth Phase**

In the fourth phase, a detailed comparative analysis was conducted to examine the consistency between human raters and AI-powered tools in scoring the 29 narrative essays. Advanced statistical methods were applied in this phase for a comprehensive examination of the data. The methods utilized in this phase included:

**One-Way ANOVA.** A one-way ANOVA was conducted to compare the means of the overall scores, as well as the means of the scores for each rubric criterion, across the five human raters and the five AI tools.

**Power Analysis.** A power analysis was performed to determine whether the sample size was sufficient to detect meaningful effects for the study's primary hypotheses, as mentioned previously in the Research Hypotheses section.

**Exploratory Factor Analysis (EFA).** An exploratory Factor Analysis (EFA) was used to identify the underlying factors that may influence the grading patterns of human raters and AI tools. For example, EFA might reveal that raters tend to group specific aspects of writing together, such as 'Organization' and 'Mechanics,' while others, such as 'Voice and Style' and 'Ideas and Content,' may emerge as distinct factors. This would suggest that raters evaluated certain criteria similarly, possibly due to shared grading principles or cognitive processes. AI tools, on the other hand, might show differences in how they group criteria, potentially highlighting areas where the AI models need refinement to align more closely with human evaluators.

**Grading Agreement Analysis.**

***Between-Group Analysis.*** Pearson correlation was used to compute the correlation between the same set of essays across different groups of graders.

***Within-Group Analysis.*** Intraclass Correlation (ICC) was used to evaluate the consistency of AI

raters as well as the reliability of human raters in evaluating the same set of essays. In order to gain insight into internal consistency across grading criteria, Pearson correlation was applied within groups to investigate the relationship between raters' scores on given pairs of criteria (pairwise comparison). The Intraclass Correlation Coefficients (ICC) were calculated to assess the reliability and consistency of scores assigned by human raters and AI tools. The ICC analysis focused on two key aspects: (1) the internal consistency of each group of raters (human raters and AI tools), and (2) the agreement between raters for each rubric criterion (e.g., 'Ideas and Content,' 'Organization,' etc.). The ICC analysis used was the two-way mixed-effects model, single rater type [ICC(C,1)], which is appropriate for assessing agreement between raters (Koo & Li, 2016). The interpretation of ICC values followed established benchmarks: poor reliability ($< 0.50$), moderate reliability (0.50–0.75), good reliability (0.75–0.90), and excellent reliability ($\geq 0.90$).

## Results

This section presents the study's results, which are organized to address the research objectives and hypothesis. First, an evaluation of the study's statistical power is provided to determine the adequacy of the sample size for detecting meaningful effects. Next, the results of the Exploratory Factor Analysis (EFA) are reported, followed by the results of the one-way repeated measures ANOVA, which compared the mean scores between human raters and AI tools to evaluate scoring consistency and identify areas of divergence. Then, the findings of the Intraclass Correlation Coefficients (ICC), which evaluate the reliability of scoring within and between groups, are reported. Finally, the results of Spearman's correlation coefficients are presented to assess the level of agreement between human and AI scores. This involves a pairwise comparison of grading consistency across criteria and a summary table highlighting correlations for all rubric criteria.

### Power Considerations

Three levels of effect size were considered in the analysis: small ($d = 0.2$), medium ($d = 0.3$), and large ($d = 0.5$). For small effect sizes, the current sample of 29 essays provided only 12% power for Hypothesis 1 and low power for Hypothesis 2. To achieve 80% power, 199 essays would have been required for Hypothesis 1 and 392 for Hypothesis 2.

For medium effect sizes, the study was underpowered with 21% power for Hypothesis 2, but it was adequately powered for large effect sizes (37 essays for Hypothesis 1). The current sample size of 29 essays was sufficiently powered to detect large effect sizes (46% power for Hypothesis 1 and 47.8% power for Hypothesis 2). In order to achieve 80% power for large effects, the study would have required 64 essays for Hypothesis 1 and 63 essays for Hypothesis 2 (see Appendix C for detailed analyses).

### Model Fit: Modeling the Dimensionality of Writing Assessment

Exploratory Factor Analysis (EFA) was used to examine the latent dimensions of the five rubric criteria: 'Ideas and Content,' 'Organization,' 'Vocabulary,' 'Voice and Style,' and 'Mechanics and Formatting.' More specifically, the analysis aimed to determine whether these criteria grouped into a single factor representing overall writing quality or reflected different dimensions of writing quality. Table 1 presents the analysis comparing the fit of a single-factor model to a two-factor model.

| Model | $\chi^2$ (df) | CFI | SRMR | RMSEA (90% CI) | Decision |
|---|---|---|---|---|---|
| Single-Factor EFA | 77.03 (5) | 0.785 | 0.110 | 0.223 (0.181–0.269) | Reject |
| Two-Factor EFA | 37.06 (3) | 0.913 | 0.076 | 0.080 (0.062–0.099) | Accept |

Table 1: Model Fit Indices for Single-Factor and Two-Factor Solutions

As shown in Table 1, the two-factor model demonstrated significantly better fit indices than the single-factor model. The two-factor model yielded a $\chi^2$ (3) value of 37.06, with acceptable fit indices: Comparative Fit Index (CFI) = 0.913, Standardized Root Mean Square Residual (SRMR) = 0.076, and Root Mean Square Error of Approximation (RMSEA) = 0.080 (90% CI: 0.062–0.099). In contrast, the single-factor model showed poor fit, with a $\chi^2$ (5) value of 77.03, CFI = 0.785, SRMR = 0.110, and RMSEA = 0.223 (90% CI: 0.181–0.269).

These findings suggested that the subjective elements, such as 'Ideas and Content' and 'Voice and Style,' were related to Factor 1, which was labeled 'Content Quality.' Conversely, the objective elements, including 'Mechanics and Formatting' and 'Organization,' were associated with Factor 2, labeled 'Mechanics and Structure.' Accordingly, Factor 1 is more connected to creativity, coherence, and tone, whereas Factor 2 is more associated with technical accuracy and structural organization.

These findings indicated that the two factors are distinct, are interrelated, and play a role in writing quality (see Appendix B for Model Fit Analysis).

**Analysis of the Mean Difference Between Human Graders and AI Tools**

The initial phase of the analysis aimed to assess grading consistency between the primary grader groups (human and AI). To achieve this, a one-way repeated measures ANOVA was conducted to explore whether the mean scores assigned by AI tools significantly differ from those given by human graders. The independent variable in this study was the grader type (human vs. AI), and the dependent variable was the scores they awarded. Table 2 presents the one-way ANOVA results, highlighting the comparative analysis between human raters and AI tools across various grading criteria.

| **Rubric Criterion** | **F-statistic** | **P-value** |
|---|---|---|
| **Ideas and Content** | 10.82 | $1.96 \times 10^{-14}$ |
| **Organization** | 4.68 | $8.62 \times 10^{-6}$ |
| **Vocabulary** | 10.36 | $8.29 \times 10^{-14}$ |
| **Voice and Style** | 13.46 | $6.54 \times 10^{-18}$ |
| **Mechanics and Formatting** | 5.56 | $4.61 \times 10^{-7}$ |

Table 2: One-Way Repeated-Measures ANOVA

As shown in Table 2, the F-statistic for the criteria 'Voice and Style' was 13.46, and the F-statistic for the criteria 'Ideas and Content' was 10.82, indicating significant differences between groups. These F-statistics represented the variance ratio between AI and humans across the groups relative to the variance within each group. High F values showed that the grades assigned by the two groups for these specific criteria could be statistically discernible. This indicated that AI and human raters may apply different standards to subjective elements such as tone and idea development.

For the 'Vocabulary' criterion, an F-statistic of 10.36 showed that AI tools and human experts analyzed word choice using potentially divergent benchmarks. This reflected varying perspectives on effective vocabulary within a context. AI tools may rely on fixed linguistic metrics, while human graders could evaluate vocabulary based on a more contextual, holistic approach, considering how well it aligns with the essay's tone or purpose.

Conversely, criteria such as 'Mechanics and Formatting' and 'Organization' show smaller, yet significant, F values of 5.56 and 4.68, respectively ($p < .01$ for both). These results suggested that while differences in scoring between AI and human raters existed for these criteria, they aligned more closely than in other categories. This alignment could be due to the more objective nature of assessing mechanical formatting and structural organization, which involved less subjective interpretation. The substantial F values for 'Voice and Style,' 'Ideas and Content,' and 'Vocabulary' confirmed that the scoring differences between AI and human graders were significant and could not be attributable to chance.

## Grading Agre]ement Analysis

### Between-Group Analysis

**Spearman's Correlation Coefficients (Monotonic Relationship).** The correlation analysis demonstrated varying levels of agreement between human and AI scores across rubric criteria, interpreted using standard correlation thresholds. Table 3 illustrates the Spearman's r correlation coefficients between human and AI scores for each rubric criterion, along with corresponding average scores and their interpretations. Greater agreement is indicated by a higher Spearman's correlation value, which implies that the two groups assessed the same rubric criteria in a comparable manner. A lower value, on the other hand, indicates disparities in grading methodologies, indicating notable variations in the evaluation of the essays by AI tools and human raters. The Spearman's correlation coefficients, corresponding average scores, and p-values between AI tools and human graders are shown in Table 3.

| Rubric Aspect | Human Scores (Avg.) | AI Scores (Avg.) | Correlation (r) | p-Value | Interpretation |
|---|---|---|---|---|---|
| Ideas and Content | 4.2 | 4.1 | 0.39 | $p < 0.001$ | Weak correlation |
| Organization | 3.8 | 3.6 | 0.52 | $p < 0.001$ | Moderate correlation |
| Vocabulary | 3.9 | 3.7 | 0.19 | $p < 0.001$ | Very weak correlation |
| Voice and Style | 3.7 | 3.6 | 0.35 | $p = 0.014$ | Weak correlation |
| Mechanics and Formatting | 3.7 | 3.6 | 0.38 | $p < 0.001$ | Weak correlation |

Table 3: Correlation Between Human and AI Scores Across Writing Rubric Aspects

Note: Correlation Strength: Very weak (0.00–0.20), Weak (0.20–0.40), Moderate (0.40–0.60), Strong (0.60–0.80), Very strong (0.80–1.00). All p-values < 0.05 indicate statistically significant correlations.

As shown in Table 3, the criteria 'Ideas and Content' had a weak positive correlation ($r = 0.39$, $p < 0.001$) and close average scores (4.2 for humans, 4.1 for AI), which indicated limited alignment in evaluating nuanced aspects such as originality and depth. The criteria 'Organization' showed the strongest correlation ($r = 0.52$, $p < 0.001$), reflecting the AI tools' relative strength in assessing structural coherence (average scores: 3.8 for humans, 3.6 for AI). In contrast, 'Vocabulary' exhibited a very weak correlation ($r = 0.19$, $p < 0.001$), highlighting the AI tools' challenges in evaluating word choice and contextual appropriateness (scores: 3.9 for humans, 3.7 for AI). The criteria 'Voice and Style' also showed weak alignment ($r = 0.35$, $p = 0.014$), indicating difficulties in capturing tone and stylistic consistency (scores: 3.7 for humans, 3.6 for AI). For 'Mechanics and Formatting,' the weak correlation ($r = 0.38$, $p < 0.001$) suggested modest agreement, with similar average scores (3.7 for humans, 3.6 for AI), reflecting that the AI tools focused on surface-level issues rather than nuanced contextual considerations.

**Within Groups Analysis**

**Intraclass Correlation Coefficients (ICC).** The Intraclass Correlation Coefficient (ICC) analysis explored how the five rubric criteria correlate within the grading of both humans and AI raters. It looked at how the different rubric criteria correlate with each other when graded by both humans and AI, aiming to detect if there was any systematic relationship between different aspects of writing (e.g., does a high score in 'Ideas and Content' correlate with a high score in 'Voice and Style'?). Table 4 presents the results of Intraclass Correlation Coefficients (ICC).

| Rubric Aspect | Rater | ICC (C,1) | 95% CI | F-Statistic (df1, df2) | Reliability | p-Value |
|---|---|---|---|---|---|---|
| **Ideas and Content** | Human | 0.55 | 0.385 – 0.716 | F(28, 112) = 7.11 | Moderate Reliability | p < 0.001 |
| | AI | 0.0917 | -0.026 – 0.273 | F(28, 112) = 1.50 | Poor Reliability | p = 0.0700 |
| **Organization** | Human | 0.585 | 0.424 – 0.742 | F(28, 112) = 8.04 | Moderate Reliability | p < 0.001 |
| | AI | 0.0519 | -0.055 – 0.223 | F(28, 112) = 1.27 | Poor Reliability | p = 0.1880 |
| **Vocabulary** | Human | 0.589 | 0.321 – 0.668 | F(28, 112) = 5.78 | Moderate Reliability | p < 0.001 |
| | AI | 0.0583 | -0.057 – 0.218 | F(28, 112) = 1.25 | Poor Reliability | p = 0.2030 |
| **Voice and Style** | Human | 0.548 | 0.280 – 0.635 | F(28, 112) = 5.07 | Moderate Reliability | p < 0.001 |
| | AI | 0.141 | 0.012 – 0.332 | F(28, 112) = 1.82 | Poor Reliability | p = 0.0148 |
| **Mechanics and Formatting** | Human | 0.504 | 0.237 – 0.597 | F(28, 112) = 4.39 | Moderate Reliability | p < 0.001 |
| | AI | 0.138 | 0.009 – 0.328 | F(28, 112) = 1.80 | Poor Reliability | p = 0.0168 |

Table 4: Intraclass Correlation Coefficients (ICC) for Human Raters and AI Tools Across Rubric Criteria

Note: Reliability interpretation based on standard ICC benchmarks: Poor reliability (< 0.50),

moderate reliability (0.50–0.75), good reliability (0.75–0.90), and excellent reliability (≥ 0.90). P-values less than 0.05 indicate statistically significant results.

According to the ICC results in Table 4, AI tools showed poor reliability across all grading criteria, especially in the subjective criteria of the rubric: Voice and Style and Ideas and Content. It is important to note that ICC values below 0.50 were generally seen as evidence of inadequate agreement among raters. In this context, the ICC values for 'Ideas and Content' and 'Organization' were recorded at 0.0917 and 0.0519, respectively, which clearly indicates a level of poor reliability. The ICC results for human raters in comparison for 'Organization' (ICC = 0.585) and 'Vocabulary' (ICC = 0.589) demonstrated moderate reliability, reflecting a higher degree of agreement among human graders who had the advantage of understanding the context and subjective nuances of the writing.

Pairwise Comparison of Grading Consistency Across Criteria Between Human and AI Graders. In the previous ICC results, an examination of the reliability between raters found that AI tools showed poor reliability across all grading criteria, especially in the subjective criteria of the rubric, while human raters showed more alignment and consistency in their grading across the given criteria. Next, Pearson correlation was applied to measure the linear relationship between pairs of criteria (e.g., the relationship between 'Ideas and Content' and 'Organization' in terms of scoring). Specifically, for a fixed essay and fixed criteria, the goal of the pairwise comparison is to compare the consistency of grading between a human grader and an AI grader. Specifically, how much agreement (or disagreement) there is between the two graders in their evaluation of the same essay was investigated using the same criteria.

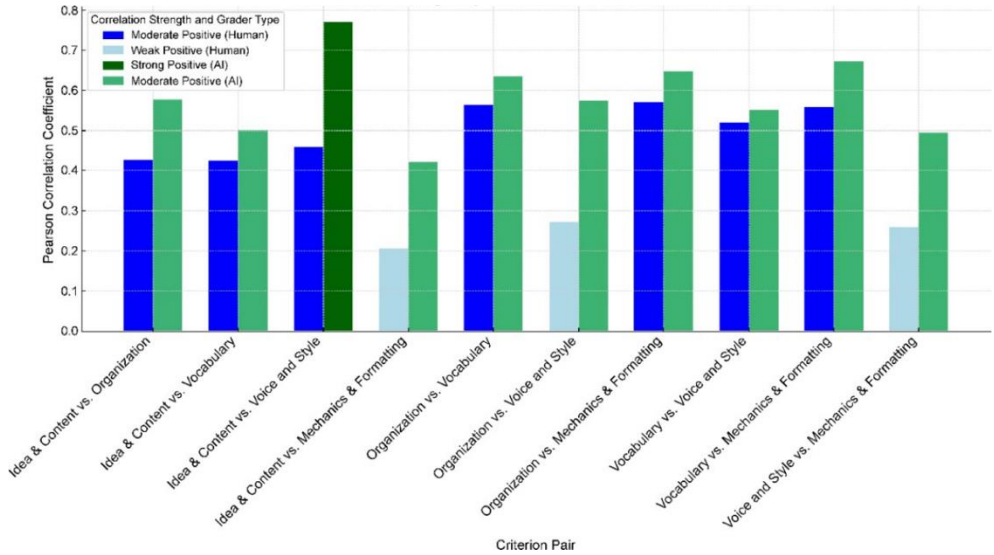


Figure 1: Correlation Strength by Criterion Pair: Human vs. AI Graders

Figure 1 illustrates the pairwise correlations between the five scoring criteria as evaluated by both human graders and AI graders, with scores aggregated by category. These correlations are quantified using the Pearson correlation coefficient. The bars in the figure are color-coded to indicate different levels of correlation strength with study-defined cut-off points as follows:

- Strong positive correlation (dark green) with a predefined range of $r \geq 0.6$ represents a high

level of agreement between criterion pairs.

- A moderate positive correlation (dark blue and light green) with a predefined range of $0.3 \leq r < 0.6$ reflects a fair level of agreement between the criteria in the pair.
- Weak Positive Correlation (light blue) with a predefined range of $0.1 \leq r < 0.3$ indicates a low but present level of agreement between the pairs.

As shown in Figure 1, the criterion 'Voice and Style' showed a correlation of $r > 0.8$ for AI models (dark green), nearly double the correlation coefficient for expert graders. This suggests that AI tools tended to score narrators more consistently by linking writing components. For instance, a high score in the 'Ideas and Content' criterion might be systematically associated with a high score in 'Voice and Style.' In other words, AI tools could overfit certain narratives by associating well-developed ideas with specific writing styles. Although this connection may seem intuitive—stronger ideas and content often relate to a clearer writing style—it is not a straightforward link, raising concerns about AI tools blending evaluative categories. This may result in less precise evaluation than those made by human graders, a topic that is explored further in the discussion section.

Moreover, human graders showed moderate to weak correlations across pairs, with $r < 0.55$ across all pairs. The highest association found for expert graders is between organization (i.e., how the Introduction, Body, and Conclusion are structured, along with transitions within the essay) and formatting (i.e., font, size, line spacing, and overall appearance of the essay). For human graders, organization and formatting often naturally influence each other. Good formatting would enhance the perceived structure of an essay, helping human graders recognize its organizational flow more easily. Conversely, poor formatting might distort the organization, making it harder to follow the writer's logic. Thus, scores for both 'Organization' and 'Mechanics and Formatting' tend to correlate, even if the actual quality varies.

## Discussion

The results of the current study detected measurable differences between AI and human scoring, particularly in subjective areas such as 'Voice and Style.' These findings aligned with prior research suggesting that AI tools, while reliable in technical tasks, struggle to fully capture the individuality of a writer's style (Magni, Park, & Chao, 2024). For objective criteria such as 'Mechanics and Formatting,' AI tools showed more consistency, supporting its potential for rule-based tasks (Rahman et al., 2023; Nazari et al., 2021). However, challenges arose in evaluating subjective criteria, evidenced by lower agreement rates between AI tools and human scores on 'Voice and Style.' Unlike human graders who could interpret context and appreciate stylistic diversity, AI tools were often confined to recognizing patterns within their training data, making them less adept at understanding complex expressions (Malik et al., 2023). These results suggest that while AI tools were helpful for preliminary assessments—especially regarding technical feedback—creative aspects still benefit from human oversight.

AI tools' strength in evaluating more objective aspects, such as 'Mechanics and Formatting,' reinforces its potential role in providing immediate rule-based feedback. This finding aligns with Rahman et al.'s (2023) finding which indicated AI tools' proficiency in detecting grammatical and structural errors with precision. Indeed, these findings support the use of AI tools for formative assessments, where timely feedback on technical elements can enhance students' learning experiences (Nazari et al., 2021). However, for more nuanced evaluations—especially those related to creativity and personal expression—human oversight remains critical (Fagbohun et al., 2024).

Interestingly, the results of this study also revealed strong correlations between certain criteria in AI tools' evaluations, particularly between 'Ideas and Content' and 'Voice and Style,' as demonstrated by the high correlations in Figure 1. This tendency suggests that AI tools may overfit certain patterns based on their training data, raising concerns regarding assessment bias (Shofiah et al., 2023). ESL students, whose writing styles reflect diverse cultural and linguistic norms, are particularly vulnerable to this bias. AI tools trained predominantly on data representing specific linguistic patterns may inadvertently disadvantage students who deviate from these norms. This highlights the ethical implications of relying solely on AI for subjective assessments, as it risks marginalizing diverse voices and perpetuating stereotypes.

Given these limitations, a hybrid approach to student assessment that combines the efficiency of AI with the nuanced judgment of human experts is recommended. As noted in prior studies, AI tools can effectively handle initial evaluations and provide rapid feedback on technical aspects (Magni et al., 2024). Still, human graders are essential for final assessments, where creativity and individual expression are key. Involving human oversight ensures that assessments are comprehensive and sensitive to the unique qualities of each student's work, thus addressing the fairness concerns associated with AI grading (Shofiah et al., 2023).

Calibration is essential for improving the reliability and fairness of AI tools. Expanding training datasets to include more diverse linguistic and cultural examples can help mitigate biases, allowing AI to adapt better to a broader range of writing styles (Malik et al., 2023). Additionally, refining AI algorithms to capture subjective elements such as creativity and voice will enhance their ability to align more closely with human evaluations. Educators also play a role in enhancing their understanding of AI tools by ensuring that they are used as supportive rather than determinative elements in the grading process (Golan et al., 2023).

Moreover, the exploratory factor analysis (EFA) demonstrated that a single-factor model could not adequately capture the multidimensional framework of writing quality, while the two-factor model distinguishing 'Content' quality and 'Mechanics and Formatting' capture the data more effectively. Recent studies highlight a significant issue with AI tools, which often fail to appropriately assess subjective writing elements such as creativity and stylistic nuance. This trend is due to the inherent biases in AI training datasets, which tend to favor dominant linguistic styles and cultural norms (Zhang et al., 2023; Peeters et al., 2021). As a result, non-native English speakers often face unfair disadvantages not because their ideas lack value, but because their writing styles do not align with the learned patterns these AI systems are programmed to recognize. This inconsistency in AI performance, especially in assessing 'Voice and Style,' emphasizes the need for systems that account for linguistic and cultural diversity in writing assessments. The training process for AI tools needs to be broadened to incorporate datasets that reflect a greater diversity of linguistic and cultural backgrounds. This expansion is essential to reducing the biases introduced by standardized training algorithms. Furthermore, hybrid models combining AI efficiency with human oversight have demonstrated promise in addressing these biases and enhancing fairness in subjective assessments (Romadhoan, 2024). In sum, refining AI systems to better evaluate subjective writing elements and linguistic diversity is a critical step toward creating equitable and culturally sensitive tools.

The poor reliability in AI tools' performance can be attributed to several underlying factors that need further exploration. While AI models such as ChatGPT, Gemini, and Claude are trained on extensive language corpora, these models often face challenges when tasked with evaluating subjective aspects of writing that require nuanced interpretive judgment. These challenges are particularly pronounced in evaluating criteria such as creativity, tone, and contextual relevance,

which are more dependent on the writer's intent and cultural context.

A closer look at the architecture of these AI models reveals that they are predominantly trained on Western-language corpora, which could create a bias toward Western writing norms. This training bias can lead to difficulty in assessing writing that reflects non-Western linguistic structures or alternative cultural perspectives, especially in creative writing. Furthermore, AI models are optimized for objective tasks, such as detecting grammar errors or structure, but they are less adept at grading more subjective components such as creativity and personal style which require deeper interpretive skills and a broader understanding of context. For example, critical elements such as tone, personality, and intent which are key to assessing 'Voice and Style' are particularly challenging for AI tools to interpret fully (Anthropic, 2023).

These limitations are amplified in subjective assessments, where context and human interpretive judgment are essential to ensure fairness and accuracy. The tool-specific training process of AI tools play a key role in basing the scores given the narrative to be textual inputs. For example, ChatGPT (OpenAI) and Claude (Anthropic) are trained to focus on user safety (OpenAI, 2023; Anthropic, 2023). This means that these tools prioritize safe outputs over strict adherence to the rubric criteria. On the other hand, Gemini (Google DeepMind) focuses on technical precision, which helps it perform better on structured factual tasks. However, this focus limits its ability to evaluate subjective aspects such as creativity and specifically emotional tone.

One of the study's key concerns was AI's potential for bias, particularly as models tend to favor patterns in their training data. This can inadvertently disadvantage students from diverse backgrounds, raising questions about fairness and inclusivity. The study recommends a hybrid assessment model that leverages both AI and human expertise, allowing for a more comprehensive evaluation that values individual expression and creativity. Future studies with more diverse data would be valuable in identifying less obvious patterns and improving AI's role in education (Golan et al., 2023). By maintaining a balance between the use of AI and human evaluation, educators can create an assessment environment that values diverse student voices and ensures fair and reliable outcomes for learners.

## Conclusions

The findings revealed both the potential and limitations of AI tools in assessing ESL students' narrative writing. While tools such as ChatGPT, Gemini, Claude, Justdone, and Chatsonic provided fast and objective feedback on technical aspects, they fell short in evaluating subjective elements such as 'Voice and Style.' This finding resonates with previous research suggesting that while AI tools are efficient, they cannot yet match the contextual understanding and empathy of human graders (Magni et al., 2024; Malik et al., 2023). The current study highlighted a significant concern regarding the potential for bias in AI tools, which often prioritize trends in their training data. As mentioned previously, this bias could unintentionally place students from diverse backgrounds at a disadvantage, prompting critical discussions around issues of fairness and inclusivity.

The results demonstrated clear patterns in the alignment between human and AI evaluations, with objective dimensions generally showing stronger correlations compared to subjective ones. This suggests that AI tools are currently more effective in quantifying structural elements, such as 'Organization,' while struggling with interpretive and context-sensitive aspects, such as 'Voice and Style.' These findings align with prior research on AI tools' limitations in subjective assessments and emphasize the need for training datasets that better reflect linguistic and cultural diversity.

In fact, the results of this study underscored the need for a hybrid assessment approach that incorporates the efficiency and consistency of AI tools with the critical and empathetic insights of human graders. Human graders provide cultural sensitivity and contextual understanding that AI tools cannot replicate at this time. A hybrid assessment approach makes the assessments not only accurate but also reflective of each student's unique voice and perspective, thereby minimizing the risks of bias from AI tools while retaining the benefits of automated grading. This idea has been supported by Naidu and Sevnarayan (2023).

Although AI grading tools could be useful for initial screening and providing insights into the overall grades, they lack the necessary training to assess the underlying pattern. Future research might focus on examining larger, more diverse datasets to further assert these findings and to identify patterns that would provide more insights into AI performance, as well as potential biases, across a broader set of educational contexts and domains. It is important to refine AI models to reduce bias and broaden their cultural adaptability. The goal is not to replace humans' judgments but to support them, enabling a more holistic assessment process that values students' unique perspectives. This refinement would make AI tools more adaptable to diverse writing styles and cultural expressions (Fagbohun et al., 2024). Additionally, equipping educators with AI literacy skills is critical to ensure that they can effectively interpret AI feedback and integrate it meaningfully into the grading process.

Using AI responsibly in education requires continuous reflection, adaptation, and commitment to fairness and inclusivity. By adopting a collaborative approach that combines the strengths of AI tools with the insights of human graders, educators can create an assessment environment that goes beyond measuring performance. This approach can help recognize and support the unique voices and creativity of every student, aligning with recent research on AI-assisted grading (Mao et al., 2024).

One limitation of the study is the insufficient statistical power to detect weak correlations (Hypothesis 2) given the sample size of 29 essays. While the study was adequately powered to assess differences in overall scores (Hypothesis 1), the ability to detect the correlation patterns between human and AI scores across rubric criteria was constrained. Future research should increase the sample size to enhance the reliability and generalizability of findings, particularly for fine-grained relationships in subjective evaluations.

Additionally, the study assessed only five AI tools and five human raters, which limited its generalizability. To identify whether the results from this sample could be applied to larger samples, future research should test more AI tools and include data from a diversified group of subjects in multiple educational levels. Although the two-factor model is helpful to explain how the writing tests are constructed, the strength of the results and their generalizability is clearly limited because of the small sample size. Future research should evaluate and further develop the two-factor model across a wide range of educational contexts in confirmatory factor analysis matched with larger and more diverse samples.

AI algorithms need to be further developed for more accurate assessment in subjective domains such as creativity and style, which are important features of writing. Future studies should, therefore, focus on improving the accuracy of AI assessments in these domains so that AI can support human judgment according to the hypered evolution model in feedback and grading.

**Acknowledgment**

The authors extend their appreciation to the Deanship of Graduate Studies and Scientific Research at the University of Bisha for funding and supporting this research under the Promising Research Program (Waed 2023), Project No. UB-Promising-25-1445.

## References

Ali, M. A. (2023). An intervention study on the use of Artificial Intelligence in the ESL classroom: English teacher perspectives on the effectiveness of ChatGPT for personalized language learning [Dissertation]. https://urn.kb.se/resolve?urn=urn:nbn:se:mau:diva-61339

Anthropic. (2023). Claude: Advancing safe AI. Anthropic. https://www.anthropic.com/research/claude

Anuyahong, B., Rattanapong, C., and Patcha, I. (2023). Analyzing the impact of Artificial Intelligence in personalized learning and adaptive assessment in higher education. International Journal of Research and Scientific Innovation, 10(4). https://doi.org/10.51244/ijrsi.2023.10412

Babbie, E. R. (2020). Practice of social research (15th ed.). Boston: Cengage Learning.

Creswell, J. W., and Creswell, J. D. (2018). Research design: Qualitative, quantitative and mixed methods approaches (6th ed.). Los Angeles: SAGE Publications.

Fagbohun, O., Iduwe, N. P., Abdullahi, M., Ifaturoti, A., and Nwanna, O. M. (2024). Beyond traditional assessment: Exploring the impact of large language models on grading practices. Journal of Artificial Intelligence, Machine Learning and Data Science, 2(1), 1–8.

Golan, R., Reddy, R., Muthigi, A., and Ramasamy, R. (2023). Artificial Intelligence in academic writing: A paradigm-shifting technological advance. Nature Reviews Urology, 20, 1–2. https://doi.org/10.1038/s41585-023-00746-x

Holmes, W., Bialik, M., and Fadel, C. (2019). Artificial Intelligence in education: Promises and implications for teaching and learning. The Center for Curriculum Redesign.

Koo, T. K., and Li, M. Y. (June 2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine, 15(2), 155-163.

Landis, J. R., and Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159–174. https://doi.org/10.2307/2529310

Liao, H., Xiao, H., and Hu, B. (2023). Revolutionizing ESL teaching with generative Artificial Intelligence—Take ChatGPT as an example. International Journal of New Developments in Education, 5(20), 39–46.

Magni, F., Park, J., and Chao, M. M. (2024). Humans as creativity gatekeepers: Are we biased against AI creativity? Journal of Business and Psychology, 39(3), 643–656.

Mahapatra, S. (2024). Impact of ChatGPT on ESL students' academic writing skills: A mixed methods intervention study. Smart Learning Environments, 11(1), 9.

Malik, A. R., Pratiwi, Y., Andajani, K., Numertayasa, I. W., Suharti, S., and Darwis, A. (2023). Exploring Artificial Intelligence in academic essay: Higher education student's perspective. International Journal of Educational Research Open, 5, 100296.

Mao, J., Chen, B., and Liu, J. C. (2024). Generative Artificial Intelligence in education and its implications for assessment. TechTrends, 68(1), 58–66.

Naidu, K., and Sevnarayan, K. (2023). ChatGPT: An ever-increasing encroachment of Artificial Intelligence in online assessment in distance education. Online Journal of Communication and Media Technologies, 13(3), e202336.

Nazari, N., Shabbir, M. S., and Setiawan, R. (2021). Application of Artificial Intelligence powered digital

writing assistant in higher education: Randomized controlled trial. Heliyon, 7(5), 3-9.
OpenAI. (2023). ChatGPT: A powerful tool for language processing. OpenAI. https://www.openai.com/research/chatgpt
Peeters, M. M. M., van Diggelen, J., and Van Den Bosch, K. (2021). Hybrid collective intelligence in a human–AI society. AI and Society. Retrieved from https://www.karelvandenbosch.nl/documents/2020_Peeters_etal_AI&S_Hybrid_collective_intelligence_in_a_human%E2%80%93AI_society.pdf.
Rahman, N. A. A., Zulkornain, L. H., Che Mat, A., and Kustati, M. (2023). Assessing writing abilities using AI-powered writing evaluations. Journal of Asian Behavioural Studies, 8(24), 1–17. https://doi.org/10.21834/jabs.v8i24
Romadhoan, M. G. E. (2024). Has AI cracked the code on English proficiency assessments? A look at how AI is revolutionizing writing evaluations. Journal of Qualitative Research in Language Education. Retrieved from https://journal.indoscholar.org/index.php/jqrle/article/download/34/25.
Sánchez-Prieto, J. C., Gamazo, A., Cruz-Benito, J., Therón, R., and García-Peñalvo, F. J. (2020). AI-driven assessment of students: Current uses and research trends. In P. Zaphiris and A. Ioannou (Eds.), International conference on human-computer interaction (pp. 292–302). Springer International Publishing. https://doi.org/10.1007/978-3-030-50513-4_22
Shofiah, N., Putera, Z. F., and Solichah, N. (2023). Challenges and opportunities in the use of Artificial Intelligence in education for academic writing: A scoping review. In Conference psychology and flourishing humanity (PFH 2023) (pp. 174–193). Atlantis Press.
Sun, T. (2023). The potential use of Artificial Intelligence in ESL writing assessment: A case study of IELTS writing tasks. Irish Journal of Technology Enhanced Learning, 7(2). https://doi.org/10.22554/ijtel.v7i2.137
Zhang, J., Wu, Z., Ridley, R., and Huang, S. (2023). Addressing linguistic bias through a contrastive analysis of academic writing in the NLP domain. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP).
Retrieved from https://aclanthology.org/2023.emnlp-main.1042/.