Human and AI Scoring of EFL Writing: The Influence of Rubrics and Genre on Reliability

Samet Taşçı samettasci@nevsehir.edu.tr
Nevşehir Hacı Bektaş Veli University

Abstract:
This study investigates the reliability of large language models (LLMs) in assessing English as a Foreign Language (EFL) writing compared to human raters. Specifically, the performances of ChatGPT 4.0 and DeepSeek R1 were examined across three genres; argumentative, opinion, and persuasive essays, under rubric-free and rubric-based scoring conditions. Participants were 65 undergraduate ELT students at a Turkish university who produced a total of 162 essays. Two experienced human raters scored all essays, and their evaluations demonstrated near-perfect inter-rater reliability, providing a stable benchmark for comparison. The same essays were then rated by ChatGPT and DeepSeek under both scoring conditions. Statistical analyses included intraclass correlation coefficients (ICC), Pearson correlations, paired-samples t-tests, and ANOVAs. Findings revealed that rubric integration substantially improved alignment between AI and human scores, particularly for ChatGPT, which showed stronger sensitivity to rubric criteria than DeepSeek. Genre effects were also evident: opinion essays yielded the highest AI-human agreement, persuasive texts moderate alignment, and argumentative essays the weakest consistency. While both AI tools produced more centralized scores with less variability than human raters, they also exhibited risk-averse tendencies, especially without rubric guidance. The results indicate that AI-based scoring can complement, but not replace, human evaluation, especially in cognitively demanding genres. The study highlights the importance of rubric clarity, prompt design, and genre awareness in maximizing the educational value of AI-assisted writing assessment.

Keywords: ChatGPT, DeepSeek, automated writing evaluation, rubric, evaluation methodologies.

Taşçı, S. (2025). Human and AI Scoring of EFL Writing: The Influence of Rubrics and Genre on Reliability. *International Journal of Education and New Approaches*, *8*(2), 191-210. https://doi.org/10.52974/jena.1785369

Submission Date:	*16.09.2025*
Acceptance Date:	*04.11.2025*
Publication Date:	*30.12.2025*

INTRODUCTION

Writing is generally considered an important part of second language (L2) learning and acquisition and is used both as a means and a measure of linguistic and communicative competence (Hyland 2019; Manchón, 2011; Williams 2012). A valid and effective approach to assessing writing skills is essential in educational institutions, as it helps the learners and the teachers know the learners' level of achievement (Leow & Suh, 2022). Hence, in EFL settings, writing assessment is crucial in determining students' achievement and informing pedagogical decisions.

Traditionally, this task has been accomplished by humans whose expertise and subjective judgment result in varied evaluations. However, the emergence of large language models (LLMs), such as ChatGPT and DeepSeek, has rapidly reshaped the landscape of writing assessment in language education. These models are increasingly being integrated into both formative and summative evaluation processes, and it raises critical questions about reliability, construct validity, and pedagogical relevance. While traditional human rating has long been the standard for assessing English as a Foreign Language (EFL) writing, the scalability and efficiency of AI-based scoring systems have led to growing interest in their classroom and institutional use.

Human assessment has been criticized for its susceptibility to inconsistencies, rater bias, and inefficiency (McConlogue, 2012; Zhao & Huang, 2020; Zhang, 2016). Furthermore, the process is time-consuming and challenging, especially in large-scale or crowded instructional contexts (Wood & Henderson, 2010). In response to these challenges, there is increasing interest in integrating Automated Writing Evaluation (AWE) systems that aim to offer standardized and potentially unbiased scores to supplement human raters (Bond et al., 2024; Khosravi et al., 2023; Mizumoto & Eguchi, 2023; Tömen, 2022). These developments have been further accelerated by the rise of generative AI technologies, particularly ChatGPT, which has led to a surge in exploratory and comparative studies investigating the capabilities, limitations, and pedagogical implications of such systems (Bui & Barrot, 2024; Kim et al., 2024; Korkmaz & Akbıyık, 2024; Shin & Lee, 2024). Although several studies report high levels of agreement between AI and human raters (Crossley, 2020), others raise concerns about genre sensitivity, limited depth in evaluation, and the tendency of AI models to produce conservative mid-range scores (Bouziane & Bouziane, 2024; Lundgren, 2024).

Writing assessment plays a crucial role in EFL instruction by measuring learners' ability to produce coherent, grammatically accurate, and genre-appropriate texts. However, the assessment process is inherently complex, influenced by the rater's background, rubric interpretation, and understanding of task requirements. The rise of AI-generated scoring brings additional complexity to this process, particularly regarding how automated systems align with human judgment across different writing genres and evaluative criteria.

Recent studies have begun to investigate the potential of AI tools such as ChatGPT for automated writing evaluation (e.g., Bui & Barrot, 2024; Kim et al., 2024; Shin & Lee, 2024). While some findings indicate promising alignment with human judgment (Crossley, 2020), others highlight inconsistencies and genre-related limitations in AI scoring (Bouziane & Bouziane, 2024; Lundgren, 2024), particularly in tasks requiring nuanced reasoning or rhetorical awareness. Despite these developments, there remains a lack of large-scale, genre-sensitive comparisons between multiple AI scoring tools and human evaluators, particularly

under different scoring conditions such as rubric-free and rubric-based assessments. This study addresses this gap by comparing the performance of ChatGPT and DeepSeek in assessing EFL student writing across three genres.

The purpose of this study is to evaluate the scoring reliability of ChatGPT and DeepSeek compared to human raters across three genres of EFL student writing, argumentative, opinion, and persuasive, under both rubric-free and rubric-based conditions. The analysis focuses on the extent to which AI-generated scores align with those of human raters and how rubric use and genre influence this alignment. The study uses statistical measures including intraclass correlation coefficients (ICC), Pearson correlations, paired-sample t-tests, and ANOVA to examine score consistency and genre effects.

This investigation is informed by frameworks from assessment literacy (Stiggins, 1995; Weigle, 2013), rater cognition theory (Eckes, 2015), and sociocultural theory (Lantolf, 2000), which provide critical lenses for understanding the cognitive, interpretive, and contextual dynamics that differentiate human and AI rating behavior. By situating AI scoring within these theoretical perspectives, the study offers a more nuanced account of the opportunities and limitations of LLMs in educational assessment.

LITERATURE REVIEW

Human Raters and Rubric Use in EFL Writing Assessment

Writing assessment in EFL education traditionally involves one or more human graders, who evaluate the essays holistically or analytically (Huang, 2008; Zhang, 2016). Despite efforts to standardize evaluations through rubrics, human scoring remains vulnerable to subjectivity and inconsistency due to factors such as rater expertise (Barkaoui, 2010), educational background (Ahmadi Shirazi, 2019), personal rating style (Zhang, 2016), linguistic background (Crusan et al., 2016), and rater fatigue (Mahshanian & Shahnazari, 2020). For instance, Barkaoui (2010) found that novice raters tend to focus disproportionately on surface-level errors, whereas expert raters prioritize rhetorical effectiveness, resulting in divergent scoring outcomes for the same student text. Such variability undermines assessment reliability, particularly in high-stakes contexts (Zhang, 2016).

These observed discrepancies are supported by Rater Cognition Models, which suggest that raters' interpretations of rubrics and written texts are filtered through their individual experiences, beliefs, and metacognitive strategies (Eckes, 2015; Lim, 2011). Even when using standardized tools, these internal filters can lead to variation in judgment. Therefore, raters' assessment literacy is essential for ensuring fairness and scoring consistency.

To address the issue of variability, standardized rubrics have become central to EFL writing assessment. Rubrics operationalize scoring criteria, thus reducing ambiguity and increasing inter-rater reliability (Dempsey et al., 2009). Analytic rubrics, which define writing performance in graded descriptors, are particularly useful in EFL contexts because they provide learners with targeted feedback (Ragupathi & Lee, 2020). However, effective rubric implementation is resource-intensive, often requiring substantial rater training and time (Wood & Henderson, 2010). These challenges, coupled with rising class sizes in global EFL programs, have encouraged the exploration of Automated Writing Evaluation (AWE) systems as scalable alternatives (Chapelle & Douglas, 2006).

AI-Based Writing Evaluation: Promise and Limitations

AI writing assessment tools apply natural language processing (NLP) and machine learning to analyze textual features at a speed and scale beyond what human raters can consistently achieve (Hussein et al., 2019). These tools mimic conventional human evaluation by assessing written output according to established criteria and writing quality indicators. Three major benefits of AI-based assessment have been widely reported: (1) fast feedback for large cohorts (González-Calatayud et al., 2021), (2) greater scoring consistency by mitigating human fatigue and bias (Bui & Barrot, 2024), and (3) support for formative learning by enabling multiple revisions (Koltovskaia, 2020).

A growing body of research has examined the reliability and validity of AI-generated writing scores compared to human raters, yet findings remain mixed. On the one hand, Bucol and Sangkawong (2024) reported a strong positive correlation between ChatGPT and human raters in evaluating student essays, concluding that generative AI has the potential to serve as a reliable scoring tool. Similarly, Shin and Lee (2024) found a high degree of scoring reliability between GPT-based chatbots and human raters, supporting the idea of using AI in educational assessment. Geçkin et al. (2023) observed a range from low to high positive correlation between ChatGPT and five human raters, while Li et al. (2024) stated that ChatGPT demonstrated good consistency and accuracy for higher grade assessments compared to human assessment and was able to differentiate between assignments of varying quality.

On the other hand, several studies have raised concerns about AI scoring consistency and depth. Bui and Barrot (2024) found that ChatGPT's scores had poor to moderate correlation with those of experienced human raters and showed inconsistencies across repeated scoring rounds. Likewise, Kim et al. (2024) reported medium reliability of ChatGPT under two different prompting conditions, suggesting a lack of robustness. Lundgren (2024) also pointed out that although GPT-4 scores aligned with average human ratings, the model demonstrated a risk-averse tendency by clustering scores in mid-range values, and changes in the scoring prompt did not meaningfully affect its performance.

Genre and Rubric Effects in AI Evaluation

While many studies have explored general AI reliability, fewer have examined how AI performance varies across different writing genres. Argumentative writing, for instance, requires logical flow and counter argumentation, while opinion writing centers on subjective expression and stance (Lu, 2011). Kim et al. (2024) highlighted that ChatGPT neglected the overall coherence of argumentative essays, suggesting its limitations in evaluating logical structure and depth of reasoning.

Bouziane and Bouziane (2024) found that while ChatGPT produced more accurate and consistent evaluations in language mechanisms, including grammar, punctuation, sentence structure, relevance, and supporting evidence than human raters, it struggled with thematic coherence and deeper contextual understanding. They concluded that AI models, optimized primarily for syntactic analysis, lack the capacity to interpret more abstract rhetorical qualities such as logic, tone, and creativity. Similarly, Mizumoto and Eguchi (2023) found that AI models performed well on linguistic dimensions but had limited ability to interpret genre-specific expectations. Lundgren (2024) confirmed that LLMs tend to focus on surface-level attributes and rarely adapt to nuanced, rubric-based standards of writing quality.

Another variable influencing AI scoring is rubric integration. Bucol and Sangkawong (2024) demonstrated that rubric-guided AI evaluations improved scoring transparency and inter-rater agreement. However, Bui and Barrot (2024) warned that AI systems may apply rubrics too rigidly, ignoring the nuanced judgments that experienced human raters apply when interpreting writing quality in context. Thus, AI scoring systems may benefit from structured criteria but still fall short in emulating human-level evaluative flexibility.

Research Gap and Study Rationale

While prior studies have provided valuable insights into the potential and limitations of AI in writing assessment, several critical gaps remain that limit our understanding of how large language models (LLMs) function as evaluative tools in authentic educational contexts.

First, few studies offer large-scale, comparative analyses of multiple AI tools evaluated under both rubric-free and rubric-based conditions. Such comparisons are essential because AI models differ in architecture and reasoning mechanisms, which may influence scoring reliability and bias. Examining them side by side under controlled conditions can reveal how design variations and prompt structures shape their evaluative behavior, thereby informing both pedagogical and technical improvement.

Second, genre variation, a core determinant of writing performance and cognitive demand, has received limited attention in AI evaluation research. Different genres (e.g., opinion, argumentative, persuasive) elicit distinct rhetorical, lexical, and organizational skills; understanding how AI models handle these variations is crucial for determining whether their scoring patterns are genre-sensitive or genre-blind. Addressing this gap enhances the validity of AI-assisted assessment across diverse communicative tasks.

Third, limited research has analyzed AI scoring reliability using human raters as benchmarks within theoretically grounded frameworks such as assessment literacy and rater cognition theory. Integrating these frameworks allows researchers to interpret AI scoring not merely as algorithmic output but as a form of "rater cognition" shaped by rubrics, task features, and prompt design. Investigating this alignment is vital for establishing AI's role in fair, interpretable, and pedagogically meaningful assessment practices.

Collectively, addressing these gaps contributes to a more comprehensive and theoretically informed understanding of how AI can complement, rather than replace, human judgment in writing evaluation. These gaps underscore the need for an integrated, empirical study that investigates how AI tools perform across genre, rubric, and comparison contexts. Therefore, the present study aims to assess the accuracy and reliability of two LLM-based AI tools (ChatGPT and DeepSeek) in scoring EFL writing tasks compared to human raters. Specifically, it examines the role of rubric use in influencing AI performance and explores how genre variation affects scoring consistency.

To fulfill the purpose of the study, the following research questions were formulated:

1. How reliable are the scores assigned by ChatGPT and DeepSeek compared to human ratings in evaluating EFL writing tasks?
2. Does providing a rubric significantly influence the accuracy and consistency of scores assigned by ChatGPT and DeepSeek?

3. How does rater agreement vary across different essay genres, namely opinion, argumentative, and persuasive, among ChatGPT, DeepSeek, and human raters?

METHOD

Research Design

This study adopted a quasi-experimental, comparative research design to evaluate the scoring performance of Generative artificial intelligence (GenAI) tools, ChatGPT 4.0 and DeepSeek R1, in comparison with human raters in assessing EFL student writing. The research focused on three distinct essay genres: opinion, argumentative, and persuasive, allowing for an investigation of genre-based variability in AI reliability. These three genres were selected because they represent the most pedagogically central and cognitively comparable forms of academic discourse taught in EFL writing curricula. They allow for systematic comparison of AI and human ratings across progressively complex rhetorical demands while maintaining structural and evaluative consistency. In addition to genre-based variability, the study compared AI scoring under two conditions: rubric-free and rubric-based evaluation. This design made it possible to assess not only the overall agreement between AI and human scoring but also the extent to which rubric integration and genre influence AI-generated assessments.

Participants and Materials

The participants in this study were 65 second-year undergraduate students enrolled in the English Language Teaching (ELT) program at a state university in Türkiye. All participants were taking a mandatory course titled Critical Reading and Writing, which aims to enhance students' academic writing proficiency, particularly in relation to coherence, critical thinking, and genre-specific rhetorical structures.

Each participant completed three writing assignments, corresponding to the following genres: opinion, argumentative, and persuasive essays. The essay topics were designed to elicit genre-appropriate responses while promoting critical engagement with socially relevant issues. For instance, the opinion essay asked students to express their views on the effects of video games, the argumentative essay focused on the dangers of technology, and the persuasive essay required students to argue for or religious education in state schools. All essays were expected to be between 250 and 300 words, a length deemed appropriate for assessing content development, coherence, and linguistic competence within a controlled format.

To ensure data integrity, only students who submitted all three genre-based assignments were included in the final dataset. This resulted in a total of 162 essays (54 per genre), written by a homogeneous group of B2-level English users. The participants' language proficiency level was determined based on their prior standardized English assessment scores. This homogeneity was intentionally maintained to reduce variability in writing quality and isolate the effects of assessment tools rather than learner differences.

All essays were submitted through the university's learning management system and anonymized prior to evaluation to ensure confidentiality. The balanced distribution of texts across genres allowed for robust cross-genre and cross-rater comparison, strengthening the internal validity of the study.

Human Rating Procedure

To establish a benchmark for comparison, all essays were independently evaluated by two experienced human raters. The first rater was a senior instructor at the School of Foreign Languages with over 15 years of teaching and assessment experience. The second rater, an assistant professor in English Language Teaching, also possessed more than 15 years of experience in academia and L2 writing instruction.

Both raters initially assessed the first 27 essays collaboratively using the provided rubric to establish rating alignment and scoring norms. After this calibration phase, the remaining essays were scored independently. A 10-point discrepancy threshold was adopted as it represents a moderate yet meaningful deviation on a 100-point scale, aligning with prior L2 writing assessment research (e.g., Cumming et al., 2002; Knoch, 2011). This criterion balances practical sensitivity with reliability control, small variations below 10 points typically reflect acceptable rater variation, whereas larger gaps may signal inconsistent interpretation of rubric criteria. Therefore, discrepancies exceeding 10 points were resolved through consensus to enhance interrater reliability and ensure the validity of the final human benchmark scores. For all analyses, the average of the two raters' scores was used as the final human rating.

To evaluate the consistency between raters, an intraclass correlation coefficient (ICC) was calculated using a two-way mixed-effects model with absolute agreement. The results indicated excellent inter-rater reliability, $ICC(3,1) = .95$, 95% CI [.93, .96], $F(161, 161) = 41.49$, $p < .001$. When the ratings were averaged, reliability increased further, $ICC(3,k) = .98$, 95% CI [.97, .98], suggesting near-perfect agreement. These reliability coefficients confirm that the human ratings could serve as a stable baseline for evaluating AI scoring performance.

AI Rating Procedure

The same set of 162 essays was evaluated by two large language model (LLM)–based AI tools: ChatGPT 4.0 (OpenAI) and DeepSeek R1. Each essay was assessed under two distinct conditions: rubric-free and rubric-based evaluation. This dual approach was intended to examine the impact of structured scoring criteria on the performance and consistency of AI-generated scores.

ChatGPT 4.0 and DeepSeek R1 were selected for this study based on their growing prevalence in educational technology applications, accessibility to educators and researchers, and proven capabilities in natural language understanding. ChatGPT has been widely integrated into writing support tools, formative feedback systems, and institutional pilot studies. DeepSeek R1, a state-of-the-art Chinese-developed LLM, offers strong multilingual capabilities and is increasingly positioned as a competitor in the global AI education space. Including both tools allows for a meaningful cross-system comparison between Western- and non-Western-developed models, thereby broadening the scope of AI performance evaluation in EFL contexts.

In the rubric-free condition, essays were submitted to ChatGPT 4.0 and DeepSeek R1 using general prompts that did not reference any scoring guidelines. The aim was to assess how well the AI tools evaluated writing based solely on their internal linguistic and algorithmic models. Both tools received the same standardized prompt text, which was carefully crafted to ensure prompt parity across systems and minimize variation caused by instruction differences. Figure 1 displays an example of the rubric-free prompt used for the argumentative essay.

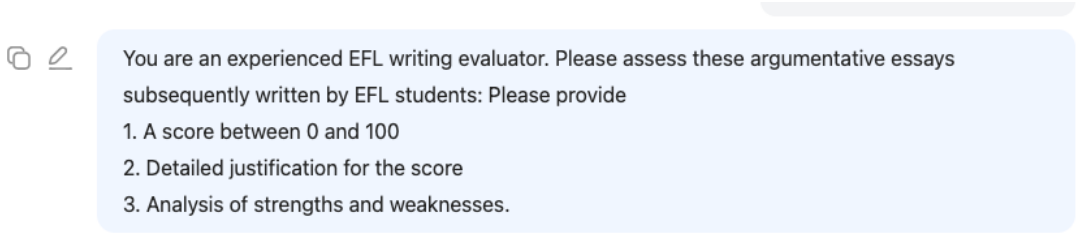

Figure 1. The prompt used for ChatGPT 4.0 and DeepSeek R1 rubric-free argumentative essay assessment.

In the rubric-based condition, the essays were re-assessed by both AI tools using a rubric-integrated prompt. This prompt explicitly included the four scoring dimensions and their weightings from the EPT Writing Rating Scale. This setup allowed for an analysis of how rubric exposure influences the scoring behavior of LLMs. The same rubric text was embedded in the prompts given to both ChatGPT and DeepSeek, and no additional stylistic or formatting cues were used. Figure 2 shows an example of the rubric-based prompt used for the opinion essay.

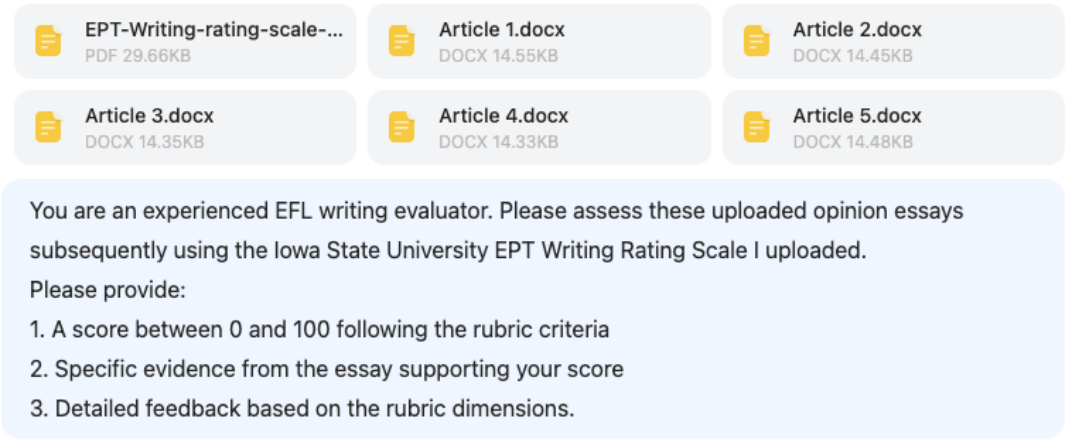

Figure 2. The prompt used for ChatGPT 4.0 and DeepSeek R1 rubric-based opinion essay assessment.

All prompts were manually entered into the AI systems by the researcher using consistent formatting, genre labeling, and instructions. For each essay, the AI's raw score output was recorded and stored for further analysis. This procedure enabled a systematic comparison of rubric-informed and rubric-independent AI evaluations and allowed for a cross-model performance comparison between ChatGPT and DeepSeek.

Rubric

The rubric used in this study was adapted from the English Placement Test (EPT) Writing Rating Scale, a widely recognized instrument for evaluating academic writing proficiency in EFL contexts. It is a holistic-analytic hybrid rubric that requires evaluators to consider four distinct scoring dimensions with designated weightings: Organization (30%), Arguments and Details (25%), Grammar and Lexis (30%), and Conventions (15%). Organization assesses the logical structure and coherence of the essay, including paragraph unity, transitions, topic development, and relevance of supporting ideas. Arguments and Details evaluate the strength, clarity, and development of claims, as well as the appropriateness and sufficiency of supporting evidence. Grammar and Lexis examine the writer's control over grammatical accuracy,

vocabulary range, and syntactic variation. Conventions include mechanics such as spelling, punctuation, paraphrasing, and formatting, as well as adherence to academic writing norms.

Each dimension is rated according to three performance levels, ranging from inadequate to competent to strong, with detailed descriptors guiding rater judgments. These descriptors were provided to both human raters and embedded within the rubric-based AI prompt to ensure scoring consistency across conditions.

Prior to the main scoring phase, both raters participated in a norming session using a sample of 27 essays representative of different proficiency levels. During this calibration process, the raters discussed each dimension in detail, aligned their interpretations of performance descriptors, and practiced assigning scores until high consistency was achieved. Both raters were already familiar with the EPT Writing Rating Scale, as it had been regularly employed in placement and proficiency assessments at their institution. Therefore, the rubric was not entirely new to them but was recontextualized for this research to ensure standardized application across human and AI evaluations.

This rubric was selected for its clarity, analytic depth, and previous validation in placement and proficiency testing contexts. Its structure enabled a transparent evaluation of genre-specific writing performance while supporting both human and AI-based assessments.

Data Analysis

The data analysis aimed to evaluate the scoring reliability and consistency of two AI tools (ChatGPT 4.0 and DeepSeek R1) compared to human raters, with particular focus on the effects of rubric usage and essay genre.

First, descriptive statistics, including mean, standard deviation, and score range, were computed for each rater group (ChatGPT, DeepSeek, and human raters) to provide an overview of the scoring patterns. Then, to evaluate overall reliability across all rater types, an Intraclass Correlation Coefficient (ICC) was calculated to assess the degree of agreement between the three rater groups: ChatGPT, DeepSeek, and the averaged human scores. A two-way mixed-effects model with absolute agreement was employed, enabling the comparison of both individual and mean ratings across human and AI evaluators. This analysis provided insight into the extent to which AI-generated scores align with human ratings under varying assessment conditions. In addition, Pearson correlation coefficients were calculated for each genre under both rubric-free and rubric-based conditions to evaluate the alignment between AI-generated and human-assigned scores. This allowed for an examination of score convergence across tools and conditions.

To assess whether rubric use significantly influenced AI scoring behavior, paired-sample t-tests were conducted for both ChatGPT and DeepSeek scores, with and without the rubric, within each genre. These analyses tested whether structured criteria impacted the consistency and accuracy of AI evaluations.

Finally, to determine whether essay genre influenced rater agreement and score variation, a One-Way ANOVA was conducted across the three genres (opinion, argumentative, and persuasive) for each assessment condition. Where significant effects were found, Tukey's HSD

post hoc tests were applied to identify specific genre-related differences in AI-human alignment.

Ethics Committee Approval

This research was conducted with the permission granted by the Nevşehir Hacı Bektaş Veli University Scientific Research and Publication Ethics Committee, based on the decision dated 05/02/2025 and numbered 2025.01.42.

RESULTS

This section presents the descriptive statistics for scores assigned by ChatGPT 4.0 and DeepSeek R1, both with and without rubric conditions, as well as two human raters and their averaged scores. These evaluations were conducted across three genres: argumentative, opinion, and persuasive essays.

As summarized in Table 1, AI-generated scores generally exhibited lower variability (i.e., smaller standard deviations) compared to human raters. This pattern suggests that AI models produced more consistently centralized scores, whereas human assessments were more dispersed. In all three genres, AI-generated scores tended to be slightly higher on average than human ratings.

In argumentative essays, DeepSeek without rubric (DSk) produced the highest mean score (M = 75.74), while the average human rating was lower (M = 69.19) and more variable (SD = 17.76). In the opinion genre, ChatGPT with rubric (GPT2) achieved the highest mean score (M = 82.48), showing a possible effect of rubric alignment. Again, human ratings showed lower means and larger standard deviations, indicating greater subjectivity or inconsistency. In persuasive essays, ChatGPT again yielded the highest average (M = 82.50), while human evaluations showed comparable means but with considerably higher variability.

These findings indicate a consistent trend: rubric-informed AI scoring, especially via ChatGPT, produced more centralized and higher average scores than human raters, while rubric-free conditions generated greater score spread, particularly among human evaluators.

Table 1. Combined descriptive statistics for all genres (N=54)

Genre	GPT Mean (SD)	GPT2 Mean (SD)	DSk Mean (SD)	DSk2 Mean (SD)	Human1 Mean (SD)	Human2 Mean (SD)	Human Avg (HA) Mean (SD)
Arg.	73.20 (6.42)	73.15 (4.78)	75.74 (6.70)	73.13 (6.72)	68.43 (18.60)	69.96 (17.35)	69.19 (17.76)
Op.	75.15 (6.77)	82.48 (5.01)	74.33 (6.14)	76.09 (7.68)	69.24 (16.08)	70.78 (14.67)	70.10 (15.30)
Pers.	74.24 (6.74)	82.50 (5.13)	75.33 (6.38)	72.91 (8.09)	71.33 (15.22)	71.39 (15.02)	71.38 (14.99)

Note: Means and standard deviations (in parentheses) are provided. GPT = ChatGPT without rubric; GPT2 = ChatGPT with rubric; DSk = DeepSeek without rubric; DSk2 = DeepSeek with rubric; HA = Human Average (mean of H1 and H2).

To assess the consistency of scoring across human and AI raters, Intraclass Correlation Coefficients (ICC) were calculated for each essay genre using a two-way mixed-effects model with an absolute agreement definition. Both single-rater reliability and average-rater reliability

were examined to evaluate the performance of individual raters and the reliability of aggregated scores.

In the argumentative genre, the reliability among individual raters was fair, ICC(2,1) = .36, 95% CI [.24, .50], $p < .001$. This value suggests considerable variation in how raters initially scored the essays, meaning that their judgments were not fully aligned when considered individually. However, when the scores were averaged across raters, reliability increased substantially to a good level, ICC(2,k) = .74, 95% CI [.62, .84]. In practical terms, this indicates that combining multiple raters' judgments produced a more stable and dependable overall evaluation, reducing the influence of individual bias or scoring differences.

For opinion essays, inter-rater reliability was weaker, with single-rater ICC(2,1) = .26, 95% CI [.14, .41], $p < .001$, suggesting poor to fair agreement among raters. Averaged scores improved reliability to a moderate level, ICC(2,k) = .64, 95% CI [.45, .78].

In the persuasive genre, the pattern persisted. The single-measure ICC was .27, 95% CI [.15, .42], $p < .001$, indicating again poor to fair agreement, while averaging scores improved reliability to ICC(2,k) = .64, 95% CI [.46, .78], suggesting moderate agreement. Table 2 summarizes ICC values across genres and scoring methods

Table 2. Intraclass correlation coefficients (ICC) by genre and rater type

Genre	ICC (Single)[b]	95% CI (Single)	ICC (Average)	95% CI (Average)
Argumentative	.364[a]	[.242, .504]	.741[c]	[.615, .836]
Opinion	.263[a]	[.140, .409]	.641[c]	[.449, .776]
Persuasive	.270[a]	[.148, .415]	.640[c]	[.464, .780]

Two-way mixed effects model where people effects are random and measures effects are fixed.
a. The estimator is the same, whether the interaction effect is present or not.
b. Type A intraclass correlation coefficients using an absolute agreement definition.
c. This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.

These findings suggest that while individual scorers, whether AI or human, exhibited limited agreement, reliability improved markedly when scores were aggregated. This pattern reinforces prior recommendations advocating for ensemble or averaged scoring approaches in AI-assisted or hybrid evaluation systems to enhance reliability and reduce variance.

To examine the degree of alignment between AI-generated scores and human ratings, Pearson correlation coefficients were computed for each genre (argumentative, opinion, and persuasive) and scoring condition (rubric-free vs. rubric-based). These analyses evaluated the extent to which ChatGPT and DeepSeek approximated human evaluative behavior under varying conditions.

In the argumentative genre, rubric-based scoring by ChatGPT (GPT2) showed the strongest correlation with human ratings, $r = .619$, $p < .01$, suggesting substantial agreement when rubric guidance was applied. DeepSeek with rubric (DSk2) showed a moderate correlation, $r = .386$, $p < .01$. Under rubric-free conditions, ChatGPT (GPT) achieved a moderate-to-strong correlation with human scores ($r = .562$), while DeepSeek (DSk) followed with $r = .500$.

For opinion essays, overall correlations were lower across both models. The highest correlation was observed for ChatGPT without rubric ($r = .379$, $p < .01$), while rubric-based scoring led to

a notable decrease in correlation ($r = .211$, n.s.), indicating a possible over alignment with rubric structure at the expense of matching human judgment. DeepSeek also showed moderate alignment in the rubric-free condition ($r = .332$, $p < .05$), with rubric-based scoring offering a slight increase ($r = .277$, $p < .05$). These findings indicate that genre characteristics interact with AI evaluation patterns in opinion essays, where expression of stance and personal reasoning often outweigh formal structure, rubric-guided scoring may overemphasize mechanical accuracy and organization, leading to weaker alignment with human raters who naturally accommodate individuality in such writing.

In the persuasive genre, the strongest correlation was observed for DeepSeek with rubric ($r = .479$, $p < .01$), reflecting better alignment in rubric-informed evaluations. ChatGPT, both with and without rubric, showed moderate correlations ($r = .260$ and $r = .413$, respectively), while DeepSeek without rubric (DSk) performed weakest ($r = .228$, n.s.). Table 3 presents the correlation coefficients by genre and AI evaluation method.

Table 3. Pearson correlation coefficients between human and AI ratings across genres

	Genre/Evaluation method	Argumentative	Opinion	Persuasive
ChatGPT	without rubric	.562**	.379**	.413**
	with rubric	.619**	.211	.260
DeepSeek	without rubric	.500**	.332*	.228
	with rubric	.386**	.277*	.479**

Note: GPT = ChatGPT; DSk = DeepSeek; GPT2 and DSk2 = rubric-based versions. $p < .05$, $p < .01$ (two-tailed).

These results highlight that AI-human alignment varies notably by genre and scoring condition. Rubric usage generally improved agreement in argumentative and persuasive writing but had inconsistent effects in opinion essays. ChatGPT showed stronger genre sensitivity, performing best in argumentative essays when guided by rubrics. In contrast, DeepSeek demonstrated more stable, but modest, correlations across conditions. Overall, these patterns suggest that genre and rubric design critically influence how closely AI models approximate human evaluative judgment.

To assess the impact of rubric use on AI-generated scores, paired-samples t-tests were conducted for both ChatGPT and DeepSeek across the three essay genres: argumentative, opinion, and persuasive. These comparisons evaluated whether rubric-based scoring significantly altered AI performance relative to rubric-free scoring.

In the argumentative genre, no significant difference was found between ChatGPT's rubric-free and rubric-based scores, $t(53) = 0.08$, $p = .940$, suggesting that rubric guidance did not meaningfully influence its scoring behavior. In contrast, DeepSeek exhibited a significant increase in scores when using the rubric, $t(53) = 3.27$, $p = .002$, indicating improved alignment or scoring inflation when structured criteria were applied.

For opinion essays, both AI systems showed statistically significant increases in scores under rubric-based conditions. ChatGPT's scores rose markedly with rubric use, $t(53) = -8.90$, $p < .001$, while DeepSeek also showed a significant increase, $t(53) = -3.30$, $p = .002$. These results suggest that rubrics enhanced scoring consistency or elevated AI sensitivity to performance indicators in opinion writing.

In the persuasive genre, rubric use produced divergent effects across models. ChatGPT's rubric-based scores were significantly higher than rubric-free scores, $t(53) = -8.24$, $p < .001$, showing strong rubric responsiveness. However, DeepSeek's rubric-free scores were actually significantly higher than its rubric-based scores, $t(53) = 2.71$, $p = .009$, suggesting that rubric guidance may have constrained its scoring behavior or altered weighting heuristics in this genre. Table 4 summarizes the paired comparisons across conditions.

Table 4. Paired-samples t-tests scores for assessments with and without rubric

		Paired Differences			t	df	Sig. (2-tailed)
		Std. Error Mean	95% Confidence Interval of the Difference				
			Lower	Upper			
Argumentative	Pair 1 GPT - GPT2	.72944	-1.40752	1.51863	.076	53	.940
	Pair 2 DSeek – Dseek2	.79894	1.00864	4.21358	3.268	53	.002
Opinion	Pair 3 GPT - GPT2	.82416	-8.98640	-5.68027	-8.898	53	.000
	Pair 4 DSeek – Dseek2	.66739	-3.54232	-.86509	-3.302	53	.002
Persuasive	Pair 5 GPT - GPT2	1.00286	-10.27073	-6.24779	-8.236	53	.000
	Pair 6 DSeek – Dseek2	.89396	.63287	4.21898	2.714	53	.009

Negative *t* values indicate higher scores with rubric; positive values indicate higher scores without rubric. $p < .05$, $p < .01$ (two-tailed).

These findings underscore the model- and genre-specific effects of rubric integration. ChatGPT appeared highly sensitive to rubric guidance, particularly in opinion and persuasive writing, which may reflect its ability to internalize structured evaluation criteria. DeepSeek showed greater variability, performing better with rubrics in some genres (opinion, argumentative), but worse in others (persuasive). This variation highlights the importance of prompt calibration and genre-aware rubric design when deploying AI scoring systems in educational contexts.

To explore whether essay genre influenced the reliability and behavior of AI-generated scoring, One-Way ANOVA tests were conducted separately for each AI model (ChatGPT and DeepSeek), under both rubric-free and rubric-based conditions. These tests assessed whether mean score differences across argumentative, opinion, and persuasive genres were statistically significant. Tukey's HSD post-hoc tests were subsequently performed to identify specific genre pairs showing significant differences.

As shown in Table 5, ANOVA results revealed significant genre effects for rubric-based scoring, particularly with ChatGPT. The model's rubric-informed evaluations (GPT2) demonstrated highly significant score differences across genres, $F(2, 159) = 63.37$, $p < .001$. DeepSeek's rubric-based evaluations (DSk2) also showed significant genre sensitivity, $F(2, 159) = 4.19$, $p = .017$.

Table 5. One-way ANOVA results for ai scores across essay genres

AI Evaluation Method	F	p-value	Sig.
ChatGPT (without rubric)	1.159	.317	Not significant
ChatGPT (with rubric)	63.374	.001	Significant
DeepSeek (without rubric)	0.688	.504	Not significant
DeepSeek (with rubric)	4.185	.017	Significant

Note. The table presents the results of one-way ANOVA analyses comparing AI-generated scores across three genres: Argumentative, Opinion, and Persuasive. A significance level of .05 was used.

In contrast, rubric-free scoring methods yielded no statistically significant differences across genres. For ChatGPT without rubric, F = 1.16, p = .317, and for DeepSeek without rubric, F = 0.69, p = .504. These results suggest that genre effects are primarily activated when rubric guidance is embedded into the evaluation process.

Tukey's HSD post-hoc tests further clarified these findings. For ChatGPT with rubric, opinion essays received significantly higher scores than both argumentative and persuasive essays (Mean Differences = 9.33 and 9.35, $p < .001$), indicating that rubric-driven evaluations may amplify genre-specific scoring tendencies.

For DeepSeek with rubric, significant differences emerged between argumentative and opinion essays (p = .043), and between opinion and persuasive essays (p = .029). These results highlight DeepSeek's more modest but still meaningful genre sensitivity under rubric conditions.

Table 6. Tukey's HSD post-hoc comparisons for AI scores across essay genres

AI Method	Group 1	Group 2	Mean Difference	p-value	95% CI Lower	95% CI Upper	Sig.
GPT2	Argumentative	Opinion	9.33	< .001	7.07	11.6	Yes
GPT2	Argumentative	Persuasive	9.35	< .001	7.08	11.6	Yes
DSk2	Argumentative	Opinion	-3.407	0.043	-6.73	-.079	Yes
DSk2	Opinion	Persuasive	3.62	0.029	-6.95	-.302	Yes

Note. GPT2 and DSk2 refer to rubric-based evaluations. Only the comparisons with statistically significant differences are shown. Post-hoc analysis was conducted using Tukey's HSD test with $\alpha = .05$.

These results underscore the importance of genre in shaping AI scoring behavior, particularly when evaluation is guided by rubric criteria. ChatGPT showed a strong tendency to rate opinion essays more favorably under rubric conditions, while DeepSeek revealed more subtle, genre-dependent variations. Collectively, these findings suggest that genre awareness and task calibration are essential when integrating AI into educational writing assessment frameworks.

DISCUSSION

This study examined the reliability and scoring consistency of two large language models (LLMs), ChatGPT and DeepSeek, in assessing EFL student essays across three genres and under two scoring conditions (with and without rubric). The findings contribute to the expanding research base on AI-assisted writing assessment by offering empirical insights into the reliability of AI-generated scores, the influence of rubric use, and the genre-based reliability in AI scoring performance. These results are interpreted in relation to prior research and their implications for assessment practice are discussed below.

Reliability of AI Scores Compared to Human Ratings

The results indicated moderate to high reliability between AI-generated scores and human-assigned ratings, particularly under rubric-based conditions. Rubric integration yielded higher intraclass correlation coefficients (ICCs) and closer alignment with human evaluations than rubric-free assessments. These findings align with Yavuz et al. (2024), who reported high ICCs (up to .972) for rubric-based scoring using fine-tuned ChatGPT models. Similarly, Shin and

Lee (2024) found strong agreement between ChatGPT and human raters when scoring tasks were clearly structured.

In contrast, our findings diverge from those of Jackaria et al. (2024), who reported low consistency between GPT-3.5 and human ratings in the absence of prompt calibration. This discrepancy may stem from our use of ChatGPT 4.0 and carefully prompt-engineered, rubric-embedded instructions, which enhanced the models' interpretive alignment with human evaluation standards. Likewise, Manning et al. (2025) found that ChatGPT performance varied across complex academic tasks, especially those requiring nuanced rhetorical evaluation, consistent with our genre-based results.

In rubric-free settings, AI models displayed risk-averse scoring behavior, clustering scores around the mid-range. This aligns with Lundgren (2024), who observed that GPT-4 assigned mid-range scores more frequently, resulting in low inter-rater reliability (Cohen's $\kappa = 0.18$). Similarly, our rubric-free scores exhibited lower alignment and reduced variability in discriminating writing quality.

The Effect of Rubric on AI Performance

The study provides robust evidence that rubric integration substantially enhances AI scoring accuracy and reliability. Across both models, rubric-based conditions produced stronger correlations with human ratings, narrower standard deviations, and clearer genre-based score distinctions. This supports findings from Bucol and Sangkawong (2024), who concluded that rubric-informed AI assessments improved score transparency and alignment. Likewise, Li et al. (2024) demonstrated that rubric-guided ChatGPT scoring improved its ability to generate detailed and justifiable assessments across writing genres.

The effectiveness of rubric use appears closely tied to prompt design and rubric clarity. Mizumoto and Eguchi (2023) emphasized that AI reliability improves when linguistic features are explicitly mapped onto rubric components. Our prompts embedded weighted dimensions and genre-specific scoring criteria, likely increasing rubric salience for the models.

These results contrast with Lundgren's (2024) findings, where minor rubric alterations did not impact GPT-4's scoring behavior. One explanation may be that Lundgren's political science rubric lacked sufficient specificity or weight distribution, making it less interpretable to the AI system. Our results suggest that rubric quality, prompt engineering, and task calibration play a crucial role in ensuring effective AI scoring performance.

The Effect of Genre on AI Performance

A central finding of this study was that AI-human scoring alignment varied significantly across genres. Agreement was highest in opinion essays, moderate in persuasive writing, and lowest in argumentative tasks, particularly under rubric-free conditions. This trend mirrors the findings of Kim et al. (2024), who reported that ChatGPT struggled with argument coherence and source-based reasoning in complex writing assignments.

Bouziane and Bouziane (2024) similarly found that while ChatGPT performed well in assessing grammar, coherence, and cohesion, it underperformed in identifying deeper rhetorical features

like thematic development and argumentative progression. Our results reinforce this limitation, particularly in argumentative texts, where cognitive complexity and counter argumentation posed challenges for both AI tools.

The persuasive genre, which blends emotional and logical appeals, yielded more moderate agreement. LLMs appeared capable of capturing surface coherence and fluency, but less effective at interpreting subtle rhetorical intent or emotional nuance. Yue (2024) also observed that ChatGPT 4.0 could recognize basic structural flaws in persuasive writing but faltered when evaluating argument strength and stylistic depth.

Genre effects may also be shaped by training corpus biases. Most LLMs are trained on general web-based content, which emphasizes narrative, expository, or transactional genres, while offering less exposure to academic argumentation or genre-specific conventions. As a result, AI tools may lack the schematic knowledge required to assess conventions critical to argumentative writing such as thesis structure, counterarguments, and logic chains.

CONCLUSION, IMPLICATIONS, AND FUTURE DIRECTIONS

This study compared the scoring performance of two large language models, ChatGPT and DeepSeek, against experienced human raters in evaluating EFL writing across three genres under rubric-free and rubric-based conditions. The results revealed distinct model- and genre-specific patterns: ChatGPT aligned most closely with human ratings in argumentative essays, whereas DeepSeek performed better with persuasive writing. Rubric integration generally enhanced scoring consistency but did not uniformly improve agreement across all genres, particularly in opinion essays where flexible reasoning and personal stance appeared to challenge AI precision.

Taken together, these findings make two key contributions. First, they provide comparative evidence on model-specific reliability in L2 writing assessment, offering one of the earliest systematic contrasts between ChatGPT and DeepSeek. Second, they illuminate how genre and rubric structure interact with AI evaluation, highlighting the contextual limits of automated scoring. Rather than positioning AI as a replacement for human raters, this study conceptualizes large language models as supportive instruments for reliable, scalable, and criterion-referenced writing evaluation.

At the classroom level, AI tools can enhance writing instruction by generating formative feedback, reducing teacher workload, and promoting students' self-assessment literacy. However, their use should be guided by clear, weighted rubrics embedded in prompts to ensure consistency. Teachers should be aware that in genres requiring higher-order reasoning, such as argumentative writing, AI may misinterpret nuance or rhetorical intent, underscoring the need for human oversight in complex evaluative contexts. Incorporating AI assessment literacy into teacher education programs will help educators design effective prompts, interpret AI feedback critically, and apply results ethically.

At the institutional level, AI scoring systems can serve as calibration tools to harmonize inter-rater consistency and facilitate norming sessions. Educational institutions adopting such systems should establish transparent assessment frameworks that ensure fairness, data privacy, and alignment with curricular objectives. Policies should promote hybrid scoring models,

combining the efficiency of AI with the interpretive expertise of human raters, especially in high-stakes or placement contexts.

Despite its robust design and comprehensive dataset, this study is subject to several limitations. The sample was limited to a single EFL population at one Turkish university, potentially constraining generalizability across contexts and proficiency levels. Furthermore, only two AI models were tested; future studies should expand to include additional LLMs such as Claude, Gemini, or open-source systems. The study also focused primarily on numerical score reliability; future research should include qualitative analyses of AI-generated feedback to assess its pedagogical appropriateness. This study did not manipulate or compare prompt engineering strategies, which are known to influence LLM outputs. Future investigations should examine how different prompt formats, levels of rubric detail, or genre-specific instructions affect AI scoring accuracy across learner profiles. Longitudinal research is also needed to evaluate the sustained impact of AI-generated feedback on student writing development, engagement, and metacognitive awareness in real classroom settings.

Acknowledgments

We are grateful to the students who participated in this study and to Instructor Uğur Ünalır for his invaluable assistance in evaluating the student essays.

REFERENCES

Ahmadi Shirazi, M. (2019). For a greater good: Bias analysis in writing assessment. *Sage Open*, *9*(1), 1-14. https://doi.org/10.1177/2158244018822377

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly*, *7*(1), 54-74. https://doi.org/10.1080/15434300903464418

Bond, M., Khosravi, H., De Laat, M., Bergdahl, N., Negrea, V., Oxley, E., Pham, P., Chong, S. W., & Siemens, G. (2024). A meta systematic review of artificial intelligence in higher education: a call for increased ethics, collaboration, and rigour. *International Journal of Educational Technology in Higher Education*, *21*(1). https://doi.org/10.1186/s41239-023-00436-z

Bouziane, K., & Bouziane, A. (2024). AI versus human effectiveness in essay evaluation. *Discover Education*, *3*(1), 201. https://doi.org/10.1007/s44217-024-00320-6

Bucol, J. L., & Sangkawong, N. (2024). Exploring ChatGPT as a writing assessment tool. *Innovations in Education and Teaching International*, 1-16. https://doi.org/10.1080/14703297.2024.2363901

Bui, N. M., & Barrot, J. S. (2024). ChatGPT as an automated essay scoring tool in the writing classrooms: how it compares with human scoring. *Education and Information Technologies,* 1-18. https://doi.org/10.1007/s10639-024-12891-w

Chapelle, C. A., & Douglas, D. (2006). *Assessing language through computer technology*. Cambridge University Press.

Crossley, S. (2020). Linguistic features in writing quality and development: An overview. *Journal of Writing Research*, *11*(3), 415-443. https://doi.org/10.17239/jowr-2020.11.03.01

Crusan, D., Plakans, L., & Gebril, A. (2016). Writing assessment literacy: Surveying second language teachers' knowledge, beliefs, and practices. *Assessing Writing, 28*, 43-56. https://doi.org/10.1016/j.asw.2016.03.001

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal*, *86*(1), 67-96. https://doi.org/10.1111/1540-4781.00137

Dempsey, M. S., PytlikZillig, L. M., & Bruning, R. H. (2009). Helping preservice teachers learn to assess writing: Practice and feedback in a Web-based environment. *Assessing Writing*, *14*(1), 38-61. https://doi.org/10.1016/j.asw.2008.12.003

Eckes, T. (2015). *Introduction to many-facet Rasch measurement*. Peter Lang.

Geçkin, V., Kızıltaş, E., & Çınar, Ç. (2023). Assessing second-language academic writing: AI vs. Human raters. *Journal of Educational Technology & Online Learning,* *6*(4), 1096-1108. https://doi.org/10.31681/jetol.1336599

González-Calatayud, V., Prendes-Espinosa, P., & Roig-Vila, R. (2021). Artificial intelligence for student assessment: A systematic review. *Applied Sciences*, *11*(12), 5467, 1-15. https://doi.org/10.3390/app11125467

Huang, J. (2008). How accurate are ESL students' holistic writing scores on large-scale assessments?-A generalizability theory approach. *Assessing Writing*, *13*(3), 201-218. https://doi.org/10.1016/j.asw.2008.10.002

Hussein, M. A., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: A literature review. *PeerJ Computer Science, 5*, e208. https://doi.org/10.7717/peerj-cs.208

Hyland, K. (2019). *Second language writing*. Cambridge University Press. https://doi.org/10.1017/9781108635547

Jackaria, P. M., Hajan, B. H., & Mastul, A. H. (2024). A Comparative Analysis of the Rating of College Students' Essays by ChatGPT versus Human Raters. *International Journal of Learning Teaching and Educational Research*, *23*(2), 478-492. https://doi.org/10.26803/ijlter.23.2.23

Khosravi, H., Viberg, O., Kovanovic, V., & Ferguson, R. (2023). Generative AI and learning analytics. *Journal of Learning Analytics, 10*(3), 1-6. https://doi.org/10.18608/jla.2023.8333

Kim, H., Baghestani, Sh., Yin, Sh., Karatay, Y., Kurt, S., Beck, J., & Karatay, L. (2024). ChatGPT for writing evaluation: Examining the accuracy and reliability of AI-generated scores compared to human raters. In C. A. Chapelle, G. H. Beckett, & J. Ranalli (Eds.), *Exploring artificial intelligence in applied linguistics* (pp. 73-95). Iowa State University Digital Press. https://doi.org/10.31274/isudp.2024.154.06

Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from?. *Assessing writing*, *16*(2), 81-96. https://doi.org/10.1016/j.asw.2011.02.003

Koltovskaia, S. (2020). Student engagement with automated written corrective feedback (AWCF) provided by Grammarly: A multiple case study. *Assessing Writing*, *44*, 100450. https://doi.org/10.1016/j.asw.2020.100450

Korkmaz, H., & Akbıyık, M. (2024). Unlocking the potential: Attitudes of tertiary level EFL learners towards using AI in language learning. *Participatory Educational Research, 11*(6), 1-19. https://doi.org/10.17275/per.24.76.11.6

Lantolf, J. (Ed.) (2000). *Sociocultural theory and second language learning*. Oxford University Press.

Leow, R. P., & Suh, B-R. (2022). Theoretical perspectives on writing, corrective feedback, and language learning in individual writing conditions. In R. M. Manchón & C. Polio (Eds.), *Routledge handbook of second language acquisition and writing* (pp. 9-21). Routledge. https://doi.org/10.4324/9780429199691-3

Li, J., Jangamreddy, N. K., Hisamoto, R., Bhansali, R., Dyda, A., Zaphir, L., & Glencross, M. (2024). AI-assisted marking: Functionality and limitations of ChatGPT in written assessment evaluation. *Australasian Journal of Educational Technology, 40*(4), 56-72. https://doi.org/10.14742/ajet.9463

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing*, *28*(4), 543-560. https://doi.org/10.1177/0265532211406422

Lu, X. (2011). A Corpus-Based evaluation of syntactic complexity measures as indices of College-Level ESL writers' language development. *TESOL Quarterly*, *45*(1), 36-62. https://doi.org/10.5054/tq.2011.240859

Lundgren, M. 2024. Large Language Models in Student Assessment: Comparing ChatGPT and Human Graders. *arXiv preprint arXiv:2406.16510*.

Mahshanian, A., & Shahnazari, M. (2020). The effect of raters' fatigue on scoring EFL writing tasks. *Indonesian Journal of Applied Linguistics, 10*(1), 1-13. https://doi.org/10.17509/ijal.v10i1.24956

Manchón, R. M. (2011). Writing to learn the language: Issues in theory and research. In R. M. Manchón (Ed.), *Learning-to-Write and Writing-to-Learn in an Additional Language*, (pp. 61-82). Johns Benjamins Publishing Company.

Manning, J., Baldwin, J., & Powell, N. (2025). Human versus machine: The effectiveness of ChatGPT in automated essay scoring. *Innovations in Education and Teaching International*, 1-14. https://doi.org/10.1080/14703297.2025.2469089

McConlogue, T. (2012). But is it fair? Developing students' understanding of grading complex written work through peer assessment. Assessment & Evaluation in Higher Education, 37(1), 113-123. https://doi.org/10.1080/02602938.2010.515010

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics, 2*(2), 100050. https://doi.org/10.1016/j.rmal.2023.100050

Ragupathi, K., & Lee, A. (2020). Beyond fairness and consistency in grading: The role of rubrics in higher education. In C. S. Sanger & N. W. Gleason (Eds.), *Diversity and inclusion in global higher education: Lessons from across Asia* (pp. 73–95). Palgrave Macmillan.

Shin, D., & Lee, J. H. (2024). Exploratory study on the potential of ChatGPT as a rater of second language writing. *Education and information technologies*, *29,* 24735-24757. https://doi.org/10.1007/s10639-024-12817-6

Stiggins, R. J. (1995). Assessment literacy for the 21st century. *Phi Delta Kappan, 77*(3), 238-245.

Tömen, M. (2022). Automated Essay Scoring Feedback in Foreign Language Writing: Does it coincide with instructor feedback? *Disiplinler Arası Dil Araştırmaları, 4*(4), 53-62. https://doi.org/10.48147/dada.60

Weigle, S. C. (2013). English as a second language writing and automated essay evaluation. In M. D. Shermis & J. Burstein (Eds.), *The handbook of automated essay evaluation: Current applications and new directions* (pp. 36-54). Routledge.

Williams, J. (2012). The potential role(s) of writing in second language development. *Journal of Second Language Writing, 21*, 321-331. https://doi.org/10.1016/j.jslw.2012.09.007

Wood, E. H., & Henderson, S. (2010). Large cohort assessment: depth, interaction and manageable marking. *Marketing Intelligence & Planning*, *28*(7), 898-907. https://doi.org/10.1108/02634501011086481

Yavuz, F., Çelik, Ö., & Çelik, G. Y. (2024). Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. *British Journal of Educational Technology*, *56*(1), 150-166. https://doi.org/10.1111/bjet.13494

Yue, X. (2024). A comparative study on ERNIE Bot 4.0 Turbo and ChatGPT 4O's performance in evaluating First-Year undergraduate persuasive essays. *Arts Culture and Language*, *1*(9). https://doi.org/10.61173/nk1ywa21

Zhang, J. (2016). Same text different processing? Exploring how raters' cognitive and meta-cognitive strategies influence rating accuracy in essay scoring. *Assessing Writing*, 27, 37-53. https://doi.org/10.1016/j.asw.2015.11.001

Zhao, C., & Huang, J. (2020). The impact of the scoring system of a large-scale standardized EFL writing assessment on its score variability and reliability: Implications for assessment policy makers. *Studies in Educational Evaluation*, *67*, 100911. https://doi.org/10.1016/j.stueduc.2020.100911

JENA
Journal of Education and New Approaches

Article Information:	Human and AI Scoring of EFL Writing: The Influence of Rubrics and Genre on Reliability
Article Type:	*Research Article*
Submission Date:	*16.09.2025*
Acceptance Date:	*04.11.2025*
Publication Date:	*30.12.2025*
Corresponding Author:	*Samet Taşçı, samettasci@nevsehir.edu.tr*
Review:	*Double-Blind Peer Review*
Ethical Statement:	** It is declared that scientific and ethical principles were followed during the preparation of this study, and all utilized works are cited in the references.*
Similarity Check:	*Conducted Turnitin*
Ethics Committee Approval:	*This research was conducted with the permission obtained by the decision of the Ethics Committee of Nevşehir Hacı Bektaş Veli University, dated 05/02/2025 and numbered 2025.01.42*
Participant Consent:	*Informed Consent Form was obtained from the participants.*
Financial Support:	*No financial support was received from any institution or project for this study.*
Conflict of Interest:	*There is no conflict of interest between individuals and institutions in the study.*
Author Contribution:	*-*
Copyright & License:	*The journal holds the copyright of the works published in the journal, and the works are published under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.*

Declaration of Use of Artificial Intelligence

During the preparation of this research, the authors employed AI tools such as ChatGPT for enhancing the fluency of the text, simplifying complex sentences for better clarity, explaining and justifying complicated constructs, and splitting lengthy sentences into shorter ones to ease comprehension. These applications assisted in ensuring the manuscript's readability while maintaining academic rigor. After using these AI tools, the author(s) reviewed and edited the content as necessary, taking full responsibility for the content of the publication. It's crucial to note that all data and findings come from properly cited sources, not AI-generated. The author(s) fully ensure the research's integrity and accuracy.