Journal of Theoretical Educational Sciences, 18(4), 736-754, October 2025
Kuramsal Eğitimbilim Dergisi, 18(4), 736-754, Ekim 2025
[Online]: http://dergipark.org.tr/akukeg
DOI number: http://doi.org/10.30831/akukeg.1718511

Can AI Assess Writing Skills Like a Human? A Reliability Analysis

Yapay Zekâ Yazma Becerilerini İnsan gibi Değerlendirebilir mi? Bir Güvenirlik Analizi

Hüseyin ATASEVEN[*] **Ömay ÇOKLUK-BÖKEOĞLU**[**]
Fazilet TAŞDEMİR[***]

Received: 12 June 2025 **Research Article** **Accepted:** 22 October 2025

ABSTRACT: This study investigates the reliability and consistency of a custom GPT-based scoring system in comparison to trained human raters, focusing on B1-level opinion paragraphs written by English preparatory students. Addressing the limited evidence on how AI scoring systems align with human evaluations in foreign language contexts, the study provides insights into both strengths and limitations of automated writing assessment. A total of 175 student writings were evaluated twice by human raters and twice by the AI system using analytic rubric. Findings indicate excellent agreement among human raters and high consistency across AI-generated scores, but only moderate alignment between human and AI evaluations, with the AI showing a tendency to assign higher scores and overlook off-topic content. These results suggest that while AI scoring systems offer efficiency and consistency, they still lack the interpretive depth of human judgment. The study highlights the potential of AI as a complementary tool in writing assessment, with practical implications for language testing policy and classroom pedagogy.

Keywords: Automated AI scoring, writing assessment, interrater reliability.

ÖZ: Bu çalışma, İngilizce hazırlık öğrencilerinin yazdığı B1 düzeyindeki görüş paragraflarını değerlendirmede özel olarak yapılandırılmış GPT tabanlı bir puanlama sisteminin, eğitimli insan değerlendiricilerle karşılaştırıldığında güvenirlik ve tutarlılığını incelemektedir. Yapay zekâ tabanlı puanlama sistemlerinin yabancı dil bağlamında insan değerlendirmeleriyle ne ölçüde örtüştüğüne dair sınırlı kanıtlardan yola çıkan çalışma, otomatik yazma değerlendirmesinin güçlü ve zayıf yönlerine ışık tutmaktadır. Toplam 175 öğrenci yazısı, hem iki insan değerlendirici hem de yapay zekâ sistemi tarafından, analitik bir rubrik kullanılarak iki kez puanlanmıştır. Bulgular, insan değerlendiriciler arasında mükemmel düzeyde bir uyum ve yapay zekâ puanlamaları arasında yüksek derecede tutarlılık olduğunu; ancak insan ve yapay zekâ değerlendirmeleri arasında yalnızca orta düzeyde bir uyum bulunduğunu göstermektedir. Ayrıca yapay zekânın daha yüksek puan verme ve konu dışı içerikleri gözden kaçırma eğiliminde olduğu görülmüştür. Bu sonuçlar, yapay zekâ sistemlerinin yazma değerlendirmesinde verimlilik ve tutarlılık sağlasa da insan yargısının sahip olduğu yorumlama derinliğinden yoksun olduğunu ortaya koymaktadır. Çalışma, yapay zekânın yazma değerlendirmesinde insan değerlendirmenin yerine değil, tamamlayıcısı olarak kullanılabileceğini vurgulamakta; dil testi politikaları ve sınıf içi pedagojik uygulamalar için önemli yansımalar sunmaktadır.

Anahtar kelimeler: Yapay zekâ ile puanlama otomasyonu, yazma değerlendirmesi, değerlendiriciler arası tutarlılık.

[*] *Corresponding Author*: Lecturer, Karabük University, Karabük, Türkiye, huseyinataseven@karabuk.edu.tr, https://orcid.org/0000-0001-9992-4518
** Prof. Dr., Ankara University, Ankara, Türkiye, cokluk@education.ankara.edu.tr, https://orcid.org/0000-0002-3879-9204
*** Assoc. Prof. Dr., Recep Tayyip Erdoğan University, Rize, Türkiye, fazilet.tasdemir@erdogan.edu.tr, https://orcid.org/0000-0002-0430-9094

Citation Information
Ataseven, H., Çokluk-Bökeoğlu, Ö., & Taşdemir, F. (2025). Can AI assess writing skills like a human? A reliability analysis. *Kuramsal Eğitimbilim Dergisi [Journal of Theoretical Educational Science], 18*(4), 736-754.

e-ISSN: 1308-1659

Measurement and assessment are fundamental components of education, playing a crucial role in evaluating student learning, guiding instructional decisions, and ensuring educational accountability. Assessments can be formative, designed to support learning through feedback, or summative, used to evaluate proficiency at a specific point in time (Brown, 2018). In language education, assessment becomes even more complex due to the multifaceted nature of linguistic competence, which includes listening, speaking, reading, and writing (Bachman & Palmer, 2010).

Language assessment typically includes both objective and subjective items, each serving different purposes in evaluating learners' proficiency. Objective items are structured questions with clear, predetermined correct answers, such as multiple-choice, true/false, and fill-in-the-blank questions. These items assess discrete linguistic components like grammar, vocabulary, and reading comprehension with high reliability and consistency (Brown, 2018). Since objective items yield quantifiable results, they minimize rater bias and ensure scoring uniformity across test-takers (Bachman & Palmer, 2010). In contrast, subjective items require human judgment in scoring and interpretation, as they often involve open-ended responses such as opinion-based writing tasks, essays, and speaking performances. These items assess productive language skills, which demand not only grammatical accuracy but also coherence, cohesion, argumentation, and rhetorical effectiveness (Weigle, 2002). The challenge with subjective items lies in their inherent variability; raters may interpret scoring criteria differently, leading to inconsistencies even when standardized rubrics are used (Jonsson & Svingby, 2007).

Writing assessment, in particular, presents challenges due to its subjective nature and the multiple elements that contribute to writing quality, such as coherence, cohesion, grammatical accuracy, vocabulary range, and argumentation (Weigle, 2002). In English Language Teaching (ELT), ensuring fairness, consistency, and reliability in writing assessment is a critical concern. While trained instructors rely on rubrics and descriptors to evaluate student writing, variability in human ratings—often influenced by rater bias, fatigue, or differing interpretations—remains a major issue (Jonsson & Svingby, 2007). As a response, artificial intelligence (AI) has been increasingly integrated into assessment practices, offering automated scoring systems that analyze linguistic features using natural language processing (NLP), large language models (LLM) and machine learning (ML) algorithms (Shermis & Burstein, 2013). These systems promise consistency and efficiency, yet questions remain about their ability to evaluate complex aspects of writing such as argument strength and coherence (Perelman, 2020).

Recent studies (Bui & Barrot, 2024; Geçkin et al., 2023; Tang et al., 2024; Wetzler et al., 2024) have therefore explored the integration of AI into writing assessment. As AI-powered tools advance, researchers have investigated their reliability, consistency, and alignment with human scoring practices, particularly in English as a Second Language (ESL) and English as a Foreign Language (EFL) context where both linguistic accuracy and discourse quality matter. Findings have been mixed. Some studies emphasize AI’s potential to reduce subjectivity, while others highlight weaknesses in detecting nuanced discourse features, proportional biases, or off-topic responses.

Several recent studies have directly compared AI-based scoring systems with human raters, highlighting both areas of alignment and persistent challenges. Geçkin et al. (2023) and Wetzler et al. (2024) found that AI-generated scores could approximate human evaluations to some extent, but inconsistencies remained—particularly with ChatGPT-3.5, which tended to diverge more from human judgments than newer models. Similarly, Ragolane et al. (2024) observed that rubric-guided AI scoring produced greater stability than manual grading, yet struggled with aspects such as creativity and nuanced argumentation. Bui and Barrot (2024) reported parallel concerns, noting that while AI could apply scoring criteria, its reliability over time was inconsistent. Together, these studies suggest that while AI can support writing assessment by reducing variability, it continues to fall short in evaluating deeper rhetorical and discourse-level features that human raters are better able to capture. In contrast, Li et al. (2024) and Tang et al. (2024) highlighted the potential of advanced models, such as hierarchical neural networks and explainable AI systems, to improve contextual understanding and transparency, although these approaches remain limited in capturing all dimensions of writing quality. Complementary perspectives are offered by DiSabito et al. (2024) and Wu et al. (2025), who argued that AI is most effective when guided by well-structured rubrics and used to support, rather than replace, human evaluation. Taken together, these findings indicate both the promise and the limitations of AI in achieving reliability, fairness, and validity in writing assessment.

Taken together, these studies indicate that while AI scoring systems can enhance efficiency, stability, and transparency, they continue to fall short in capturing the interpretive depth of human evaluation. Importantly, most existing research has focused on general-purpose AI models such as ChatGPT, often without rubric alignment (Hannah et al., 2023; Lundgren, 2024; Rahman et al., 2022). This leaves a critical gap: few studies have investigated custom GPT-based models aligned with analytic rubrics in specific educational contexts, such as B1-level English writing (Dazzeo, 2024; Geçkin et al., 2023; Sun, 2024). Addressing this gap is significant both theoretically, for advancing discussions on reliability and validity, and practically, for improving fairness and accuracy in foreign/second language assessment.

Accordingly, this study aims to evaluate the effectiveness of a custom GPT-based scoring system in comparison to trained human raters. The following research questions guided the investigation:

RQ1: How reliable and consistent are scores produced by a custom GPT-based system compared to trained human raters?

RQ2: To what extent do the scores of the custom GPT-based system align with those of human raters in evaluating B1-level opinion paragraphs?

RQ3: To what extent do AI-generated scores align with those assigned by human raters?

RQ4: How effectively does the AI system evaluate off-topic writing samples?

By answering these questions, the study examines whether a tailored GPT model, aligned with an analytic rubric, can serve as a reliable complement to human judgment in EFL writing assessment.

Method

Research Design

Building on the theoretical background and research gap outlined in the introduction, this section describes the methodological framework employed to investigate the reliability and validity of the custom GPT-based scoring system in comparison with human raters.

This study adopted a descriptive-analytical design (Creswell & Creswell, 2018), which included both descriptive statistics (means, standard deviations, and score distributions) to summarize the data and inferential statistics (Intraclass Correlation Coefficients, ICC) to evaluate inter-rater reliability. The design was structured to directly address the research questions. Specifically, RQ1, which examined the reliability and consistency of scores, was analyzed using ICC tests across GPT-based and human ratings. RQ2, which investigated the extent of alignment between GPT-generated and human ratings, was examined through comparisons of descriptive statistics and corresponding ICC values. ICC was selected as the primary reliability index because it is the most appropriate measure for continuous rating data involving multiple raters, whereas alternatives such as weighted kappa are generally suited for categorical scoring.

Participants

The participants in this study consisted of 175 students enrolled in the English preparatory program at a state university in Türkiye. All participants were active B1-level students in the department and represented the entire population of learners at this level during the study period. The group was homogeneous in terms of nationality, repetition status, and faculty background. Of the participants, 52% were male, with a mean age of 18 years (SD = 0.6). Their B1 level was determined by the institutional placement test, on which they had an average score of 70 (SD = 8). Most of the students were enrolled in engineering faculties, while a smaller proportion were in departments where English was used either as the full (100%) or partial (30%) medium of instruction. A convenience sampling procedure was employed, as all available students meeting the B1-level criterion were included in the study. In addition, two trained raters participated in the study. The raters were experienced English language instructors, aged 35 and 37, with 12 and 14 years of teaching experience, respectively, and had prior training in rubric-based writing assessment.

Procedures

Prior to data collection, students were informed that their written responses would be used for research purposes, and all identifiable information was anonymized to uphold ethical standards related to confidentiality and data protection. No personally sensitive data was retained or reported.

The data for this study were collected through an in-class writing task administered as part of routine assessment in the English preparatory program. Participants were instructed to compose an opinion paragraph in response to a prompt aligned with B1-level curricular objectives. The writing task was conducted under standardized conditions, ensuring consistency in timing, instructions, and topic

exposure. Each student's response was anonymized prior to evaluation to ensure objectivity and compliance with ethical research standards. The opinion paragraphs were independently scored by two experienced English instructors using an analytic scoring rubric that evaluated key dimensions such as task fulfillment, coherence and cohesion, lexical resource, and grammatical accuracy.

Data Collection

The final dataset comprises a total of 175 writing samples produced by B1-level students enrolled in the English preparatory program. Each sample was evaluated with blind grading independently by two experienced human raters and twice by the custom GPT-based AI scoring system, resulting in four scores per response. To compare human and AI scoring, each paragraph was also assessed twice by a custom GPT-based scoring system designed to align with the same rubric. This allowed for analysis of both intra-AI consistency and agreement with human scores. In addition, a set of 10 writings decided to be off-topic by human raters was specifically examined to evaluate the AI system's capacity to detect content irrelevance.

Customized GPT for Writing Assessment

The ChatGPT platform offers the capability to create customized GPTs, enabling users to tailor the model's behavior and responses to specific domains or tasks. By setting personalized instructions and uploading supporting materials, users can enhance the model's performance in specialized contexts. This adaptability supports more precise and context-aware applications across fields such as education, research, and professional development, representing a notable advancement in user-centered artificial intelligence.

For the purposes of this study, a specialized GPT model was developed to function as an automated assessment tool. This model is specifically designed to evaluate opinion paragraphs written by university students enrolled in English preparatory programs at the B1 proficiency level. Its design reflects an informed alignment with the linguistic competencies expected at this level and the academic writing standards typically upheld in higher education.

The model employs a detailed 20-point scoring rubric encompassing four key dimensions: Content (7 points), Organization & Structure (7 points), Grammar (3 points), and Vocabulary (3 points). The GPT was trained to be able to detect and respond to topic relevance; paragraphs deemed off-topic or marginally related to the prompt are evaluated using an adjusted scoring approach that accounts for reduced task fulfillment.

In addition to assessing language use and idea development, the model also incorporates specific point deductions for common writing issues such as omitting a title, misusing linking expressions, or failing to meet the required word count. This ensures that evaluations reflect not only linguistic proficiency but also adherence to fundamental academic writing conventions. The full set of instructions used to configure this GPT is provided in Appendix A.

Writing Assessment Rubric

In writing assessment, many educators argue that greater importance should be placed on the depth and organization of students' ideas rather than on superficial language elements like grammar and vocabulary. Although grammar and mechanics are often highlighted in writing instruction, research suggests they should not overshadow the evaluation of content and structure. According to Wolf and Stevens (2007), the ability to effectively convey ideas and organize them logically holds more value in the assessment process. Reflecting this perspective, the rubric used in this study consists of four main categories: content, organization & structure, grammar, and vocabulary. The rubric assigns higher weight to content and organization & structure, with 7 points each, while grammar and vocabulary are allotted 3 points each. This scoring scheme underscores the emphasis on idea development and coherence over linguistic precision alone, aligning with the construct of writing assessment as described by Brown (2018).

Data Analysis

A total of 175 unique opinion paragraphs were evaluated twice by human raters and twice by the AI system. To examine the consistency and reliability of these evaluations, the Intraclass Correlation Coefficient (ICC) was calculated using R, an open-source statistical computing platform (R Core Team, 2025). The ICC is a commonly used metric for assessing the level of agreement or consistency between raters or repeated measurements, making it especially appropriate for studies involving multiple evaluations of the same content (Koo & Li, 2016). In this study, a two-way random-effects model with absolute agreement (ICC[2,1]) was employed to assess single-measure reliability, and the corresponding average-measure ICC was also reported to capture overall consistency across raters. Reliability was interpreted based on the thresholds suggested by Portney and Watkins (2000): ICC values below 0.50 suggest poor reliability, values from 0.50 to 0.75 indicate moderate reliability, values between 0.75 and 0.90 indicate good reliability, and values above 0.90 suggest excellent reliability (Koo & Li, 2016).

Ethical Procedures

Ethics committee approval for this study was obtained from the Social Sciences and Humanities Research Ethics Board of Karabük University on March 28, 2025, under the document number E-78977401-050.04-423166.

Results

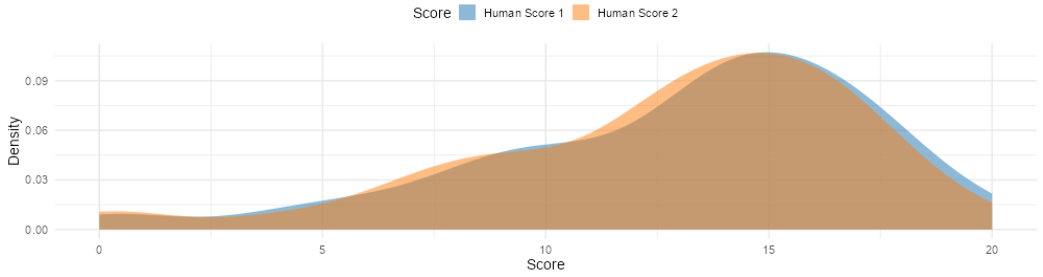
The following section presents the results of the analyses conducted to address the research questions outlined in the introduction.

Agreement between Human Raters

To gain a clearer understanding of how each human rater assigned scores to the student opinion paragraphs, the distributions of their scores were visualized. This type of visual representation helps to reveal the overall scoring tendencies of each rater, including whether scores were clustered around certain values or more widely dispersed across the scoring scale (Wickham, 2016). The distribution of scores assigned by human raters is presented in Figure 1.

Figure 1

Distribution of Human Scores Across Two Rating Sessions



As can be seen in Figure 1, the overlaid density plots illustrate the overall similarity in score patterns, with both distributions peaking around similar values, suggesting a general alignment in scoring trends. Minor differences in distribution shape and spread may reflect variation in scoring sensitivity between the two raters. To better visualize the alignment between human scores, a scatter plot is provided in Figure 2.

Figure 2

Scatter Plot of Human Scores Across Two Rating Sessions

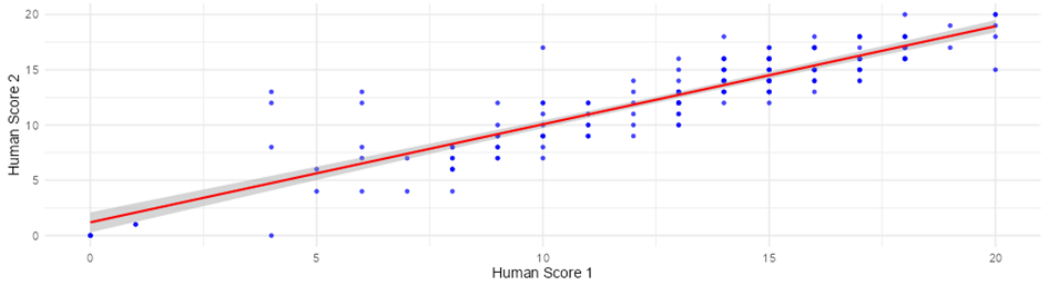


Figure 2 provides a visual representation of the relationship between the scores assigned by the two human raters. The scatter plot shows that the data points closely follow a linear pattern, indicating a strong agreement between the two sets of scores. The fitted regression line and narrow confidence interval further support this consistency, suggesting that both raters applied the scoring rubric in a reliable and uniform manner. This visual evidence aligns with the high interrater reliability observed in the statistical analysis.

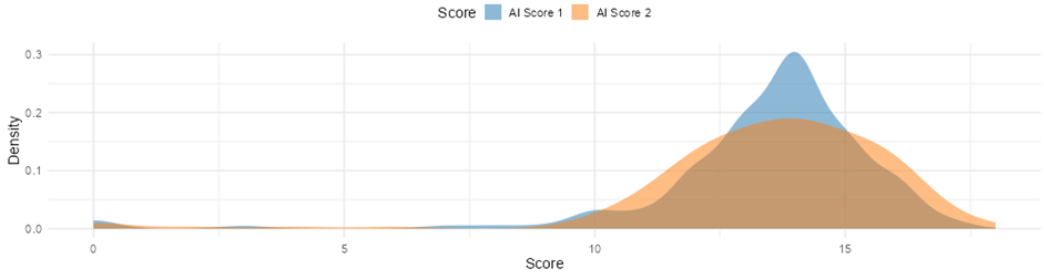
To evaluate the consistency of human raters in assessing the student writing samples, the ICC was calculated using a two-way mixed-effects model for absolute agreement and average measures. The analysis revealed an ICC(A,2) of 0.94, indicating excellent interrater reliability (Koo & Li, 2016; Portney & Watkins, 2000). This result suggests a high degree of agreement between the two human raters in evaluating the opinion paragraphs. The F-test confirmed that this level of agreement was statistically significant, $F(174, 171) = 18.1$, $p < .001$, with a 95% confidence interval ranging from 0.925 to 0.959. These findings underscore the reliability of the human scoring process and support its use as a reference point for comparing AI-based evaluations in subsequent analyses.

Consistency Across Multiple AI-generated Scores

To gain a clearer understanding of how each AI rater assigned scores to the student opinion paragraphs, we visualized the distributions of their scores. The distribution of scores assigned by AI raters is presented in Figure 3.

Figure 3

Distribution of AI Scores Across Two Rating Sessions



As shown in Figure 3, the overlaid density plots illustrate the distribution of scores generated by the AI system across two separate rating instances. While both AI Score 1 and AI Score 2 exhibit similar means, indicating general consistency in central scoring tendencies, there are notable differences in how the scores are distributed. Specifically, AI Score 1 shows a sharper peak, reflecting higher kurtosis and suggesting that the model assigned a larger number of scores clustered tightly around the mean. In contrast, AI Score 2 displays a flatter, wider distribution, indicating greater dispersion in scoring. This contrast suggests that, although the two sets of AI scores align in terms of average performance, the scoring behavior differs slightly in variability—possibly due to subtle shifts in how the model weighed certain rubric components across sessions. To better visualize the alignment between AI scores, a scatter plot is provided in Figure 4.

Figure 4

Scatter Plot of AI Scores Across Two Rating Sessions

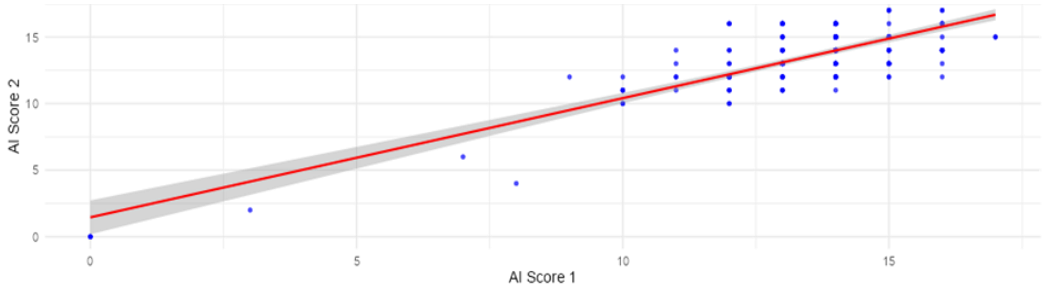


Figure 4 visualizes the consistency between two independent scoring instances performed by the AI system. Each dot represents a student's writing sample scored in two separate sessions by the custom GPT-based model. The data points closely follow the red regression line, indicating a strong positive correlation between the two AI-generated scores. The narrow gray confidence band suggests minimal variability, highlighting the internal stability of the scoring mechanism. While some dispersion is visible in the mid-to-lower score range, the overall alignment implies that the AI model applies scoring criteria in a highly consistent manner when operating under identical conditions.

However, Figure 4 shows that the majority of AI-generated scores fall above the 10-point mark, suggesting a general tendency toward lenient scoring. This pattern raises concerns about the AI system's ability to penalize off-topic or partially relevant responses effectively. Despite the model's overall consistency, its apparent reluctance to assign lower scores may indicate limitations in detecting content that diverges from the prompt, particularly in cases where surface-level linguistic features remain strong but semantic relevance is weak.

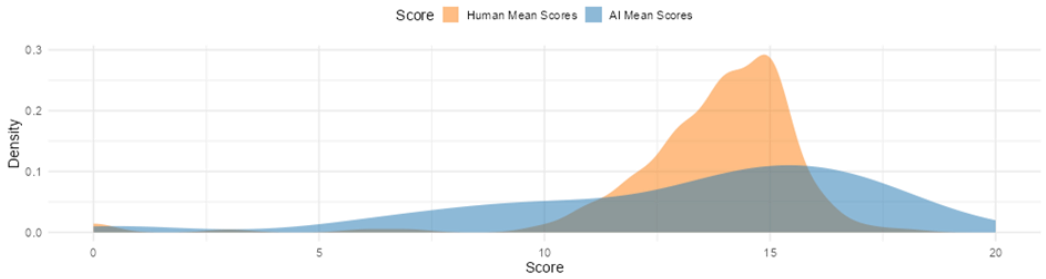
To determine the consistency of the custom GPT-based scoring system, the ICC was calculated for two separate AI-generated scores of each opinion paragraph. The analysis, based on a two-way mixed-effects model with absolute agreement for average measures, yielded an ICC(A,2) of 0.900. According to the interpretive guidelines by Koo and Li (2016), this value indicates excellent reliability. The F-test result, $F(174, 174) = 9.99$, $p < .001$, confirmed that the observed agreement was statistically significant. The 95% confidence interval ranged from 0.866 to 0.926, further supporting the internal consistency of the AI system in evaluating student writing. These findings suggest that the GPT-based scoring model produces stable results when applied multiple times under identical conditions.

Alignment of AI-Generated Scores with Human Rater Scores

To explore the alignment between human-assigned scores and AI-generated scores, the distributions of their respective mean values were plotted and are presented in Figure 5. This visual comparison helps to highlight areas of agreement as well as potential differences in how each scoring source evaluates student writing performance.

Figure 5

Distribution of Human and AI Mean Scores



As shown in Figure 5, although the means of the human and AI scores are generally similar, there are clear differences in the shapes of their distributions. The human scores (in orange) form a sharper, more peaked curve, indicating a high concentration of scores near the average. In contrast, the AI-generated scores (in blue) exhibit a broader, flatter distribution, reflecting greater variability in score assignments. This wider spread suggests that the AI system differentiates more across writing performances, while human raters may apply scoring criteria more conservatively or consistently around a central range. To better visualize the alignment between human mean scores and AI mean scores, a scatter plot is provided in Figure 6.

Figure 6

Scatter Plot of Human and AI Mean Scores

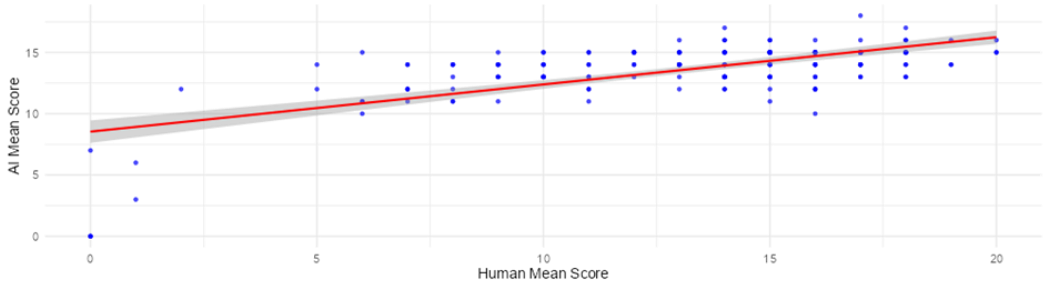


Figure 6 compares the average scores assigned by human raters and those generated by the AI system. The scatter plot reveals a positive correlation, as indicated by the upward-sloping regression line. However, a noticeable pattern emerges: AI scores tend to cluster more tightly within the mid-to-high score range, while human scores demonstrate greater variability across the full scale. The relatively flatter slope of the regression line and the presence of several outliers—particularly where human scores are low but AI scores remain moderate—suggest that the AI system may overestimate weaker performances.

This trend implies that while the AI model aligns moderately well with human judgment on higher-scoring responses, it may struggle to mirror human sensitivity to lower-quality or off-topic content. The AI's tendency to assign scores within a narrower band could reflect limitations in detecting nuanced issues such as lack of coherence, weak argumentation, or irrelevance to the prompt—elements human raters are trained to penalize.

To assess the degree of agreement between the AI-generated scores and the scores assigned by human raters, the Intraclass Correlation Coefficient (ICC) was computed using the average of two human scores and two AI scores per writing sample. The analysis employed a two-way mixed-effects model for absolute agreement. The resulting ICC(A,2) was 0.725, which, according to Koo and Li (2016), indicates moderate to good reliability. The agreement was statistically significant, $F(174, 170) = 3.69$, $p < .001$, with a 95% confidence interval ranging from 0.630 to 0.796. These findings suggest that while the AI scoring system demonstrates a fair degree of alignment with human judgment, its evaluations are not fully interchangeable with those of trained human raters.

Effectiveness of the AI System on Evaluating Off-Topic Writing Samples

To investigate the AI system's ability to detect off-topic responses, a total of 175 paragraphs were analyzed. Among these, 10 paragraphs had been independently identified as off-topic based on their clear divergence from the assigned prompt. Both human raters and the custom GPT-based scoring model were asked to classify each paragraph as on-topic, somewhat on-topic, or off-topic.

Human raters successfully identified all ten off-topic paragraphs (100% accuracy), whereas the AI system detected only four of them (40% accuracy). The AI model thus failed to recognize six off-topic writings (60% misclassification rate), labeling them as either on-topic or somewhat on-topic.

These findings suggest that the AI system struggles to reliably differentiate between nuanced levels of topic relevance. Despite being trained with task-specific instructions and a rubric that emphasized topical alignment, the model exhibited a tendency to overlook semantic drift or superficial adherence to the prompt. This raises concerns about the AI's effectiveness in detecting off-topic writing, an important factor in ensuring the validity of writing assessment.

Discussion and Conclusion

This study aimed to evaluate the reliability and alignment of a custom GPT-based scoring system against trained human raters for B1-level opinion paragraphs. The findings offer a nuanced perspective on the current capabilities and limitations of AI in writing assessment. While the results confirmed the AI's capacity for high internal consistency, they also highlighted a significant gap between its automated evaluations and the interpretive depth of human judgment, particularly concerning scoring leniency and the detection of off-topic content.

A primary finding was the moderate alignment between human and AI scores. This discrepancy appears to stem from the AI's tendency to be more lenient, especially with lower-quality writing. This pattern suggests the AI model may overvalue surface-level linguistic features, such as grammatical correctness and lexical variety, while failing to adequately penalize weaknesses in deeper discourse-level components like argumentation, coherence, and task fulfillment. Human raters, by contrast, demonstrated a more holistic evaluative approach, effectively integrating both linguistic accuracy and content quality into their scores. This aligns with the theoretical view that writing proficiency is a multifaceted construct that extends beyond mere mechanical correctness. The AI's struggle to penalize irrelevant content suggests its current algorithms are not yet sophisticated enough to fully grasp semantic context and authorial intent, which remain key strengths of human evaluation.

These results both corroborate and extend the findings of previous research in automated writing assessment. In line with the observations of Geçkin et al. (2023) and Wetzler et al. (2024), our study found that AI-generated scores often diverge from human judgments, with a notable tendency toward leniency. Similarly, our findings echo Ragolane et al. (2024), who noted that while rubric-guided AI can achieve stability, it struggles with nuanced aspects like creativity and argumentation. This study adds to this body of work by demonstrating that even a custom GPT model, specifically trained on an analytic rubric designed to prioritize content and organization, is not immune to these shortcomings. The system's pronounced failure to identify off-topic responses underscores a critical limitation that persists despite advances in model customization.

The practical implications of these findings are significant for educators and institutions considering the integration of AI into their assessment practices. The results strongly suggest that AI scoring systems, in their current form, are not suitable to be used as standalone, high-stakes evaluation tools. Their leniency and inability to reliably detect off-topic content could lead to inflated grades and invalidate assessment outcomes. Instead, the AI's strengths, consistency and efficiency position it as a valuable supplementary tool. For instance, it could be used for low-stakes formative assessments, providing students with instant feedback on grammatical and lexical

features, thereby freeing up instructors to focus on higher-order concerns such as idea development and critical thinking. This human-AI partnership model, as advocated by DiSabito et al. (2024) and Wu et al. (2025), appears to be the most pedagogically sound approach, leveraging AI's efficiency without sacrificing the validity and fairness that human oversight ensures.

In conclusion, this study confirms that while a custom GPT-based system can score student writing with a high degree of internal consistency, it does not yet fully replicate the nuanced and context-aware judgment of experienced human raters. The AI demonstrated a tendency toward lenient scoring and a significant weakness in identifying off-topic content, revealing limitations in its ability to assess deeper levels of writing quality. Therefore, AI should be viewed as a supportive instrument rather than a substitute for human evaluators in the writing assessment process. Its proper role is to enhance efficiency and provide formative feedback, but final, summative judgments—especially where fairness and validity are paramount—must remain under the purview of human educators.

Limitations and Suggestions for Future Research

Although this study provides meaningful insights into the reliability and consistency of a custom GPT-based scoring system, several limitations should be acknowledged. First, the research focused on a single proficiency level (B1) within one institutional context in Türkiye. Therefore, the findings may not be generalizable to different proficiency levels, writing genres, or educational settings. Future research could include participants from various institutions and proficiency levels to examine whether similar patterns of reliability and alignment persist across broader linguistic and contextual conditions.

Second, the customized GPT model used in this study was developed through prompt-based configuration rather than fine-tuning with large-scale annotated data. Consequently, its scoring behavior reflects guided adaptation rather than data-driven learning. Future work could explore models trained or fine-tuned on larger, rubric-aligned writing corpora to assess whether such models better capture nuanced linguistic and rhetorical features.

Third, while the Intraclass Correlation Coefficient (ICC) provided a solid measure of interrater reliability, this approach primarily captures numerical agreement and does not fully reflect the qualitative aspects of scoring. Future studies may benefit from incorporating discourse-level analysis, linguistic feature extraction, or explainable AI methods to better understand the reasoning processes behind AI-generated scores.

Fourth, the number of writing samples (N = 175) and the use of a single writing genre (opinion paragraphs) limit the breadth of inference. Writing performance and AI interpretations may differ significantly across genres such as narrative, argumentative, or expository writing.

Future research should expand the range of genres and include larger datasets to better evaluate the scalability and generalizability of AI scoring tools. Enhancing AI models to detect deeper semantic relationships and topic relevance would be critical, particularly in formative contexts where task fulfillment is as important as language quality. Additionally, the integration of hybrid scoring models—where AI offers a first-

pass evaluation followed by human review—may balance efficiency with interpretive depth and mitigate some of the observed shortcomings.

Finally, while human ratings were used as the benchmark for comparison, human scoring is itself subject to bias and interpretive variation. The study did not include adjudication procedures or qualitative calibration discussions that could further validate differences between human and AI scores. Future research could employ mixed-method approaches that combine quantitative reliability measures with qualitative analyses of rater–AI disagreements to gain a more comprehensive understanding of how judgments are formed.

In sum, while the present findings indicate that AI-based scoring systems can achieve high consistency, they also highlight the need for continued refinement in areas such as semantic understanding, rubric sensitivity, and contextual validity. Future studies that integrate AI and human expertise in complementary ways may offer the most balanced and pedagogically sound approach to writing assessment.

Acknowledgements

This study was conducted independently and did not receive any specific grant or funding from public, commercial, or not-for-profit institutions. The authors declare that the research was self-initiated and carried out without external financial support.

Statement of Responsibility

Lect. Hüseyin Ataseven contributed to the study in terms of conceptualization, design, data collection and processing, analysis and interpretation, literature review, and writing the original draft. Prof. Dr. Ömay Çokluk–Bökeoğlu contributed through supervision, conceptualization, design, interpretation, literature review, editing, and critical review. Assoc. Prof. Dr. Fazilet Taşdemir contributed to the design of the study, data analysis and interpretation, literature review, writing, and critical review.

Conflicts of Interest

The authors declare that there are no conflicts of interest related to the research, authorship, or publication of this study.

Author Bios:

Lect. Hüseyin Ataseven received his B.A. in English Language Teaching from Gazi University in 2011. Since 2012, he has been working as an English Lecturer at Karabük University, School of Foreign Languages. He completed his master's degree in Educational Measurement and Evaluation at Ankara University in 2023 and is currently pursuing his PhD in the same field at Ankara University. His academic interests focus on assessment and evaluation in language education, test development, and statistical analysis in educational research. Ataseven has contributed to various institutional projects and international conferences. In addition to his teaching responsibilities, he is actively engaged in academic research and interdisciplinary collaborations, aiming to bridge the gap between educational theory and practice, particularly in the context of foreign language instruction and assessment.

Prof. Dr. Ömay Çokluk-Bökeoğlu completed her undergraduate studies in the Guidance and Psychological Counseling program at Ankara University, Faculty of

Educational Sciences. She earned her master's degree from the university's Social Sciences Institute and completed her PhD in Educational Statistics and Research at the Institute of Educational Sciences. Since 1997, she has been a faculty member at Ankara University's Faculty of Educational Sciences. Her research interests include methodology, educational statistics, multivariate statistics, and scale development. She has authored numerous books, articles, and conference papers in these fields, contributing extensively to the academic literature. Prof. Dr. Çokluk-Bökeoğlu is recognized for her work in advancing quantitative research methods and statistical analysis in educational research.

Assoc. Prof. Dr. Fazilet Taşdemir graduated from the Faculty of Education at Karadeniz Technical University in 2004 and began her academic career the same year as a research assistant. She continued her postgraduate studies at Ankara University, earning her master's degree in 2008 with a thesis on the construct validity of the TIMSS mathematics subtest and prediction of achievement, and her PhD in 2013 with a study on classification accuracy using standard-setting methods. She currently serves as a faculty member in the Department of Educational Sciences, Guidance and Psychological Counseling Division at Recep Tayyip Erdoğan University. Dr. Taşdemir has served as an advisor on numerous theses and holds editorial and reviewer roles in various national and international journals. Her research interests include multivariate and advanced statistics, classification accuracy, test development and adaptation, item writing and scoring, and research methods. She is actively involved in academic publishing and contributes significantly to the field of educational measurement.

References

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford University Press.

Brown, H. D. (2018). *Language assessment: Principles and classroom practices* (3rd ed.). Pearson Education.

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

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches.* Sage.

Dazzeo, R. (2024). AI-enhanced writing self-assessment: Empowering student revision with AI tools. *Journal of Technology-Integrated Lessons and Teaching, 3*(2), 80–85. https://doi.org/10.13001/jtilt.v3i2.9119

DiSabito, D., Hansen, L., Mennella, T., & Rodriguez, J. (2025). Exploring the frontiers of generative AI in assessment: Is there potential for a human-AI partnership?. *New Directions for Teaching and Learning*, *2025*(182), 81-96. https://doi.org/10.1002%2Ftl.20630

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

Hannah, L., Jang, E. E., Shah, M., & Gupta, V. (2023). Validity arguments for automated essay scoring of young students' writing traits. *Language Assessment Quarterly, 20*(4–5), 399–420. https://doi.org/10.1080/15434303.2023.2288253

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity, and educational consequences. *Educational Research Review, 2*(2), 130–144. https://doi.org/10.1016/j.edurev.2007.05.002

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Li, C., Long, J., Guan, N., Wang, Y., Liang, W., & Li, C. (2024, August). AI-Assisted College Students' English Writing Scoring with Hierarchical LSTM for Enhanced Contextual Understanding and Grading Accuracy. In *2024 7th International Conference on Education, Network and Information Technology (ICENIT)* (pp. 175-180). IEEE. https://doi.org/10.1109/ICENIT61951.2024.00039

Lundgren, M. (2024). Large language models in student assessment: Comparing ChatGPT and human graders. *arXiv preprint arXiv:2406.16510.*

Perelman, L. (2020). The BABEL generator and E-rater: 21st century writing constructs and automated essay scoring (AES). *Journal of Writing Assessment*, *13*(1).

Portney, L.G., Watkins, M.P., (2000). *Foundations of clinical research: applications to practice.* New Jersey: Prentice Hall.

R Core Team. (2025). *R: A language and environment for statistical computing* (Version 4.4.1) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Ragolane, M., Patel, S., & Salikram, P. (2024). AI Versus Human Graders: Assessing the Role of Large Language Models in Higher Education. *Asian Journal of Education and Social Studies, 50 (10). pp. 244-263*. https://doi.org/10.9734/ajess/2024/v50i101616

Rahman, N. A. A., Zulkornain, L. H., & Hamzah, N. H. (2022). Exploring artificial intelligence using automated writing evaluation for writing skills. *Environment-Behaviour Proceedings Journal, 7*(SI9), 547–553. https://doi.org/10.21834/ebpj.v7iSI9.4304

Shermis, M. D., & Burstein, J. (2013). *Handbook of automated essay evaluation: Current applications and new directions*. Routledge.

Sun, Y. (2024). AI in teaching English writing: Automatic scoring and feedback system. *Applied Mathematics and Nonlinear Sciences, 9*(1), 1–17. https://doi.org/10.2478/amns-2024-3262

Tang, X., Chen, H., Lin, D., & Li, K. (2024). Incorporating Fine-Grained Linguistic Features and Explainable AI into Multi-Dimensional Automated Writing Assessment. *Applied Sciences*, *14*(10), 4182. https://doi.org/10.3390/app14104182

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis* (2nd ed.). Springer. https://doi.org/10.1007/978-3-319-24277-4

Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press.

Wetzler, E. L., Cassidy, K. S., Jones, M. J., Frazier, C. R., Korbut, N. A., Sims, C. M., & Wood, M. (2024). Grading the Graders: Comparing Generative AI and Human Assessment in Essay Evaluation. *Teaching of Psychology*, 52(3), 298-304. https://doi.org/10.1177/00986283241282696

Wolf, K., & Stevens, E. (2007). The role of rubrics in advancing and assessing student learning. *Journal of effective teaching*, *7*(1), 3-14.

Wu, X., Saraf, P. P., Lee, G., Latif, E., Liu, N., & Zhai, X. (2025). Unveiling scoring processes: Dissecting the differences between llms and human graders in automatic scoring. *Technology, Knowledge and Learning*, 1-16. https://doi.org/10.1007/s10758-025-09836-8

Appendix A

The Prompt to Create "B1 Level - Opinion Paragraph Scoring GPT"

You are a highly skilled and experienced professor with a PhD in English Language Teaching. Your primary role is to evaluate opinion paragraphs written by university students enrolled in an English preparatory class at the B1 level. You have developed a clear and structured rubric to ensure fair, valid, and reliable assessment. The writing test is graded out of 20 points.

B1 Level Writing Exam Rubric (Total: 20 points)

Criterion 1: Relevance to the Given Topic (Determines Scoring Eligibility)

1. On-topic (Eligible for Full Scoring – Out of 20): The response fully addresses the given topic. It is evaluated based on all the criteria below.

2.Somehow Off-topic (Eligible for Partial Scoring – Out of 10): The response is somewhat related to the topic but contains noticeable deviations, unclear connections, or unnecessary details. It is evaluated using a modified rubric:

Content: 0 points, Organization & Structure: 4 points max, Grammar: 3 points max, Vocabulary: 3 points max,

3. Off-topic (0 Points): The response does not address the given topic. No further evaluation is applied.

Evaluation Criteria (For On-Topic Responses – Scored Out of 20)

1. Content (7 points) – Evaluates topic relevance, depth of ideas, and justification.

6-7 points: Fully developed response with strong justification. Excellent coverage of the topic with relevant, clear, and sufficient examples.

5 points: Good coverage with relevant and appropriate ideas but may lack depth or examples.

4 points: Adequate response, but may have minor irrelevance, lack sufficient development, or include weak examples.

3 points: Inadequate coverage with weak or insufficient ideas.

2-1 points: Largely irrelevant response with minimal knowledge of the topic.

0 points: No relevance to the given topic or no development of ideas.

2. Organization & Structure (7 points) – Evaluates paragraph structure, coherence, and linking words.

6-7 points: Well-organized, coherent, and fluent paragraph. Includes a strong topic sentence, well-developed supporting ideas, and a clear conclusion. Smooth use of linking words.

5 points: Generally well-organized with logical flow but may have slight weaknesses in transitions or development.

4 points: Adequate organization with minor structural issues. Ideas flow logically but may include some unclear connections or redundancy.

3 points: Weak structure with misplaced sentences, insufficient supporting ideas, or a weak conclusion. Limited coherence.

2-1 points: Disorganized paragraph with disconnected ideas and a lack of logical flow.

0 points: Completely disorganized with no unity or coherence.

3. Grammar (3 points) – Evaluates sentence structure and correctness.

3 points: Excellent use of grammar with very few or no mistakes. Strong sentence structures appropriate for the level.

2 points: Some grammatical mistakes that slightly affect clarity but do not obscure meaning.

1 point: Frequent grammar mistakes that make understanding difficult.

0 points: No correct sentence structures, causing severe comprehension issues.

4. Vocabulary (3 points) – Assesses range, correctness, and spelling.

3 points: Wide range of vocabulary with accurate use of target words. Very few or no spelling errors.

2 points: Some vocabulary repetition and minor mistakes in word choice, form, or spelling.

1 point: Limited vocabulary use with incorrect word meanings or severe spelling mistakes.

0 points: Minimal vocabulary range, excessive repetition, and numerous spelling mistakes that obscure meaning.

Appendix A

The Prompt to Create "B1 Level - Opinion Paragraph Scoring GPT" (Continued)

Evaluation Criteria (For Somehow Off-Topic Responses – Scored Out of 10)

Content (0 points): No points awarded due to topic irrelevance.

Organization & Structure (4 points): Evaluated as per the rubric but with a reduced maximum score.

Grammar (3 points): Evaluated normally.

Vocabulary (3 points): Evaluated normally.

Point Deductions for Not Following Instructions:

Organization & Structure Deductions:

- If linkers are missing or misused, deduct 1-2 points from Organization & Structure.
- If there is no title, deduct 1 point from Organization & Structure.
- If paragraph structure is weak (e.g., no clear topic sentence, no supporting details, or no conclusion), deduct 2-3 points from Organization & Structure.

Grammar Deductions:

- If there are frequent grammatical errors that make understanding difficult, deduct 1-2 points from Grammar.
- Vocabulary Deductions:
- If there is excessive repetition of words or limited vocabulary use, deduct 1-2 points from Vocabulary.
- If spelling mistakes significantly affect readability, deduct 1-2 points from Vocabulary.

Word Count Deductions:

- 100-119 words: Deduct 1 point from each category.
- 80-99 words: Deduct 2 points from each category.
- Less than 80 words: Deduct 3 points from each category.

Output Format

Student Writing Assessment

- Final Score: X/20

- Breakdown:
- Content: X/7
- Organization & Structure: X/7
- Grammar: X/3
- Vocabulary: X/3

Instructions:

You are given either a photo of a handwritten student writing sheet or a typed paragraph. Evaluate the writing based on the provided rubric.

- Assign a score for each rubric category.
- Calculate and report the final overall score.
- Provide detailed feedback explaining the rationale behind each score and why the final score is appropriate.

Appendix B

Writing Task

Duration: 45 Minutes

Allowed Sources: No dictionaries or notes are allowed

Classroom: 10 Classrooms with an invigilator in each

Students' Background: They are familiar with opinion paragraphs due to prior writing lessons and portfolio studies.

Name Surname :

Student Number :

Class :

In modern world, which below is the **BEST** way of communication?

- Sending messages
- Calling
- Video Call (Skype, Zoom, etc.)
- Sending emails

You are going to choose **ONE** of these communication ways and write **an opinion paragraph** for a school magazine.

- What is your opinion about the statement?
- Write a well-developed opinion paragraph with a topic sentence, three supporting sentences and a strong conclusion.
- Please DO NOT forget to use appropriate **linkers** (e.g. first of all, secondly, in addition, finally etc.).
- Provide **details, evidence, and specific examples** relevant to your topic sentence.
- Use **at least 120 words.**
- Please DO NOT forget to write a good title for your paragraph.