International Journal of Assessment Tools in Education
2025, Vol. 12, No. 1, 20–32

https://doi.org/10.21449/ijate.1517994

journal homepage: https://dergipark.org.tr/en/pub/ijate *Research Article*

Artificial intelligence as an automated essay scoring tool: A focus on ChatGPT

Ahmet Can Uyar[1*], Dilek Büyükahıska[2]

[1]Sivas Cumhuriyet University, School of Foreign Languages, Department of English, Sivas, Türkiye
[2]Ondokuz Mayıs University, Faculty of Education, Department of English Language Teaching, Samsun, Türkiye

ARTICLE HISTORY

Received: July 18, 2024

Accepted: Oct. 07, 2024

Keywords:

Automated essay scoring,
Artificial intelligence,
ChatGPT,
Foreign language writing,
Writing evaluation.

Abstract: This study explores the effectiveness of using ChatGPT, an Artificial Intelligence (AI) language model, as an Automated Essay Scoring (AES) tool for grading English as a Foreign Language (EFL) learners' essays. The corpus consists of 50 essays representing various types including analysis, compare and contrast, descriptive, narrative, and opinion essays written by 10 EFL learners at the B2 level. Human raters and ChatGPT (4o mini version) scored the essays using the International English Language Testing System (IELTS) TASK 2 Writing band descriptors. Adopting a quantitative approach, the Wilcoxon signed-rank tests and Spearman correlation tests were employed to compare the scores generated, revealing a significant difference between the two methods of scoring, with human raters assigning higher scores than ChatGPT. Similarly, significant differences with varying degrees were also evident for each of the various types of essays, suggesting that the genre of the essays was not a parameter affecting the agreement between human raters and ChatGPT. After all, it was discussed that while ChatGPT shows promise as an AES tool, the observed disparities suggest that it has not reached sufficient proficiency for practical use. The study emphasizes the need for improvements in AI language models to meet the nuanced nature of essay evaluation in EFL contexts.

1. INTRODUCTION

AI has become one of the indispensable parts of our everyday lives with new tools emerging each day whose functions range from advanced interaction with humans to image creation. In this context, educational settings are unsurprisingly being embellished with such tools for the purpose of enhancing the process of teaching. Not limited to the process of teaching itself, AI tools have also started to become a subject of educational assessment. In the realm of language learning and teaching, the evaluation and the assessment of written products in the target language stand as pivotal measures of linguistic prowess. Traditionally, writing evaluation has been predominantly conducted by the course instructors, drawing upon their expertise and understanding of language nuances, context, and cultural intricacies. However, the rapid advancements in AI have introduced a revolutionary shift in this landscape. AI-powered

*CONTACT: Ahmet Can UYAR ✉ ahmetcanuyar97@gmail.com Sivas Cumhuriyet University, School of Foreign Languages, Department of English, Sivas, Türkiye

e-ISSN: 2148-7456

systems, leveraging sophisticated algorithms and machine learning models, now offer an alternative or complementary method for assessing and grading EFL learners' written products.

The idea of evaluating essay writings based on a machine algorithm dates back to the 1960s, the decade when Page (1966) proclaimed that computer-based essay scoring parallel to that of human scoring was on the horizon. Page (1966) attempted to articulate what is recognized today as AES or automated writing evaluation (AWE), which can be described as the utilization of technology to assess and rate essays. It entails employing computer programs to examine and assign scores to written pieces by considering predetermined standards like language accuracy, vocabulary depth, logical flow, sentence structure, and meaningful connection. Although this technology is becoming more advanced in assessing the meaningful discourse of written works, initial AES systems faced criticism for concentrating solely on superficial aspects while overlooking content-related characteristics, leaving them susceptible to cheating tactics by learners like padding with extra words and commas (Attali, 2013). While many AES systems have been documented to focus on analyzing surface-level linguistic and structural elements, human assessment of essays prioritizes different aspects of language use and discourse (Huang, 2014). However, along with the advancements in Natural Language Processing (NLP), Large Language Model (LLM) technologies, and related AI developments, it is probable to assert that these shortcomings started to fade, and a new era for AES has started rising.

According to Huang (2014), the rise of AES stems from two primary factors. Firstly, the immense burden on educators due to teaching demands, class instructions, and the substantial time and effort required for grading students' written work is notable, accounting for nearly 30% of their workload (Mason & Grove-Stephenson, 2002). With limited time and resources, instructors struggle to effectively assess compositions and provide feedback. Introducing AES systems into classrooms could alleviate this overwhelming workload by handling and evaluating writing assignments. Secondly, as Huang (2014) noted, AES offers a distinct advantage over human evaluation by ensuring consistency in scoring, as its criteria are programmed and executed uniformly. In contrast, human assessment is susceptible to inconsistency due to cognitive fatigue, distractions, and interruptions over time. Particularly in large-scale assessments like "Test of English as a Foreign Language" (TOEFL) or "Graduate Record Examination" (GRE), human subjectivity can lead to varying and unreliable ratings.

These factors have been discussed to be overcome by developing AES technologies. Therefore, in recent years, the utilization of AES and AWE systems has garnered significant attention within educational settings for assessing written compositions, particularly among EFL learners. Numerous studies (Almusharraf & Alotaibi, 2022; Chen & Pan, 2022; Huang, 2014; Manap *et al*., 2019; Wang & Bai, 202; Zribi & Smaoi, 20211) have explored the effectiveness of various AES and AWE systems in evaluating writing proficiency, highlighting their strengths and limitations. These studies predominantly focused on employing distinct software tools such as Criterion, Grammarly, PaperRater, and other automated assessment platforms to assess writing quality. Despite the rich literature on this subject, one notable area that remains relatively unexplored is the utilization of LLMs, such as ChatGPT, in the domain of essay grading for EFL learners. Notably, while the existing research has examined the efficacy of different software tools in automated grading, the assessment potential of ChatGPT, an advanced language generation model created by OpenAI, remains relatively uninvestigated (Bui & Barrot, 2024). Therefore, this study represents an earnest effort to bridge this gap by investigating the feasibility and effectiveness of ChatGPT in evaluating EFL learners' essays, thereby contributing to the existing literature by exploring a novel avenue in AES. In this vein, a comparison was pursued in order to reveal the (un)parallelism between the scores generated by human raters and ChatGPT for the essays written by EFL learners. The investigation was shaped by the subsequent research questions:

1. Is there a significant difference between the scores generated by human raters and ChatGPT?

2. Does the genre of the essay play a significant role in the agreement between the scores produced by human raters and ChatGPT?

1.1. Automated Essay Scoring

AES systems have gained attention in education, particularly in evaluating writing proficiency for EFL learners. Much of the research effort was built on specific AES tools which were specifically designed for this pursuit, rather than LLMs or NLP models like ChatGPT. However, many of these tools were argued to fall short of adequately corresponding to the complex nature of essay scoring. This is because these specific tools were only capable of recognizing the essays from a mechanical or rule-based perspective. In line with this, studies have revealed both the benefits and limitations of these tools. For instance, Huang (2014) examined the AES tool Criterion in an EFL context, finding a weak correlation between AES and human scoring, with Criterion often assigning higher scores. The study showed that the tool focused more on language mechanics, while human raters emphasized discourse and writing quality. Huang (2014) concluded that the AES tool Criterion, though efficient, may not capture the nuanced elements of writing as well as humans.

Similarly, Manap *et al*. (2019) compared PaperRater with human evaluation and found that this AES tool was more lenient, showing a moderate correlation with human scores. Despite its utility in providing quick feedback, PaperRater's ability to assess deeper content and relevance was limited. Zribi and Smaoi (2021) echoed this finding, noting that PaperRater assigned consistently higher scores than human raters for intermediate-level EFL learners, raising concerns about its reliability. A similar finding in that the AES tool assigned higher scores than human raters was evident in Chen and Pan (2022), who explored the effectiveness of Aim Writing in improving Chinese college students' English writing skills. While the tool helped with grammar and vocabulary, human feedback was more related to structure and organization. The study found a notable correlation between the tool and human scores, although Aim Writing consistently assigned higher scores.

On the other hand, several studies revealed that certain AES tools assigned scores lower than those of human raters, a finding that is in contrast with the body of research aforementioned. Namely, Almusharraf and Alotaibi (2022) evaluated Grammarly's effectiveness compared to human raters, focusing on writing errors in 197 EFL essays. While Grammarly detected more errors, human raters assigned higher scores overall. The study highlighted that the tool excelled at catching specific grammar issues but struggled with more complex writing aspects such as sentence flow and coherence. The authors recommended using the tool as a supplementary tool rather than a standalone solution. A similar finding in the context of high-quality essays was published by Wang and Bai (2021), who assessed the accuracy of two AES systems, Pigai and iWrite, using 486 essays from non-English majors. According to the findings, both systems agreed with each other but differed significantly from human raters, especially for high-quality essays, in which AES systems tended to score lower. This suggests that while AES systems can handle lower-quality writing better, they struggle with more advanced writing, emphasizing the need for further refinement.

While specific tools designed to serve as AES systems in the market still attract considerable attention, the shift recently geared towards the latest developments in AI. That is to say, AI-based chatbots which are becoming more and more capable of producing human-like speech day by day are the recent subjects of AES. One of these renowned tools which are in public use with its user-friendly interface is ChatGPT.

1.2. ChatGPT as an AES Tool

ChatGPT, which stands for chat generative pretrained transformer, is an AI-powered LLM tool developed by OpenAI (accessible at https://chat.openai.com). It helps computers understand and generate text that resembles human speech. Because ChatGPT was released relatively recently, studies that take it under focus in the context of AES are considerably restricted in number (Bui & Barrot, 2024).

OpenAI occasionally releases a new version of ChatGPT, which creates a panorama of different versions such as ChatGPT-3, ChatGPT-3.5, or ChatGPT-4. Besides, it is also possible for it to be used in connection to other tools or with modifications. Therefore, the existing literature presents an amalgam of studies that utilized various versions or variations of ChatGPT. For instance, Mizumoto and Eguchi (2023) examined the use of OpenAI's text-davinci-003 model, part of GPT-3.5, as an AES tool. Using the TOEFL11 corpus, which includes 12.100 essays, they assessed the model's accuracy and reliability, particularly when linguistic features like lexical diversity and syntactic complexity were added. Statistical analyses showed that including these features significantly improved essay scoring accuracy. While the ChatGPT model demonstrated a certain level of reliability, the study concluded that AI-based AES should support human raters rather than replace them. The authors also highlighted the need for further research into ethical concerns and student motivation in AI-based assessments.

Yancey *et al*. (2023) explored the use of LLMs, particularly ChatGPT-3.5 and ChatGPT-4, for AES in high-stakes English language tests for language learners. Using short essay responses from the Duolingo English Test which were scored by human raters, the study aimed to compare the performance of ChatGPT models to that of existing AWE systems and assess inter-rater agreement between the models and human raters. Results showed that ChatGPT-4, when given calibration examples, closely matched the performance of current AWE systems. However, the model's agreement with human ratings varied based on the test-takers' first language. The authors highlighted potential of ChatGPT-4 to enhance AWE systems, while stressing the need for careful consideration of fairness and ethical implications. Parker *et al*. (2023) examined the potential of ChatGPT-3 as an AWE tool in nursing education, focusing on providing formative feedback on scholarly writing. The study analyzed 42 graduate nursing students' papers, where ChatGPT-3 evaluated macro-level elements such as organization, development, and thesis clarity. The authors emphasized the importance of using well-constructed prompts to obtain useful feedback from generative AI tools. ChatGPT-3 was found to grade more strictly than a human rater, awarding only one paper a score of 3 (grade A) and giving a score of 2 (grade B) to the rest. It provided detailed feedback with suggestions for improvement in areas like evidence, organization, and mechanics. The authors highlighted the efficiency and individualized feedback that AI tools offer, promoting autonomous learning. The researchers emphasized the need for further research on using ChatGPT effectively in writing instruction and the importance of educating faculty and students on its implementation. The study concluded that ChatGPT holds promise for enhancing writing feedback in nursing education.

Guo and Wang (2024) investigated the potential of ChatGPT (the version was not explicitly stated in the study; however, the date of the data collection corresponds to ChatGPT-3) to assist EFL teachers in providing feedback on learners' argumentative essays. The study involved 50 essays written by Chinese undergraduate students with B2 to C1 English proficiency. Both ChatGPT and five Chinese EFL teachers, varying in experience and technology use, evaluated the essays, focusing on content, organization, and language. ChatGPT provided more feedback than the teachers, offering balanced attention to all aspects, while teachers varied in their focus. ChatGPT's feedback also included summaries that could help learners understand their writing globally, which might aid in revisions. However, ChatGPT occasionally provided off-task feedback, possibly due to its aspect-specific prompting, which was not observed in the teachers' feedback. The teachers generally rated ChatGPT's feedback positively, acknowledging its detailed and structured nature but also noting that its effectiveness depends on learners' ability to understand and apply it. The study's limitations included focusing only on argumentative essays, excluding student perspectives, and not accounting for the teachers' ability to provide more personalized feedback. The study concluded that while ChatGPT has the potential to support EFL writing instruction, its classroom integration requires careful consideration.

Bui and Barrot (2024) explored the use of ChatGPT-3.5 as an AES tool in writing classes. The researchers compared the scores generated by ChatGPT-3.5 with those assigned by a highly experienced human rater, who had 20 years of expertise in essay evaluation. The research analyzed 200 argumentative essays from college students, categorized by proficiency levels from A2 to B2. The results revealed that ChatGPT scores did not strongly correlate with those assigned by human raters, showing weak to moderate correlations and a lack of consistency, as indicated by low intraclass correlation coefficients. The authors suggested that the differences could be due to ChatGPT's scoring algorithm, the data it was trained on, changes in the model, and its inherent randomness. The study also highlighted some limitations, such as relying on a single human rater and not including qualitative feedback in the analysis. The researchers recommended future research that involves multiple raters, considers qualitative feedback, and investigates how student-level factors affect AES performance. Despite these challenges, the paper offers perspectives to ChatGPT's current capabilities as an AES tool and its potential for improvement.

In conclusion, while these studies provided valuable insights into the current capabilities of ChatGPT as an AES tool, they collectively highlighted several limitations and areas for further research. Certain scholars noted that future research should focus on exploring a broader range of essay types (Guo & Wang, 2024) and incorporating multiple human raters for comparison (Bui & Barrot, 2024). In parallel, the current study corresponds to these future research suggestions mentioned, incorporating different essay types and multiple human raters.

2. METHOD

2.1. Research Design

This study adopts a quantitative approach. Quantitative methodology is an approach used in scientific research that focuses on the collection and analysis of numerical data to answer research questions. It involves employing several strategies, techniques, and assumptions to examine various phenomena through the analysis of numerical patterns, enabling researchers to gather and analyze numeric data to conduct statistical analyses ranging from simple to complex, including aggregating data, revealing relationships, and making comparisons across aggregated datasets (Coghlan & Brydon-Miller, 2014). Creswell (2009) proposed that choosing the appropriate research design depends on the nature of the research questions of the study. Because this study seeks to compare the scores generated by two different parties from a statistical point of view, it utilizes a quantitative approach featuring the Wilcoxon signed-rank test and the Spearman correlation coefficient test.

2.2. Materials and Data Collection

A total of 50 EFL essays were collected from 10 learners (6 female and 4 male students) who studied at the B2-level English preparatory class at a state university in Türkiye. The participants and their essays were sampled based on the convenience sampling method. The essays were written by the learners as writing tasks in the context of a writing class without resorting to any type of AI tools. In line with the research aims, various types of essays were included in the sample. The compilation consists of 10 analysis essays, 10 compare and contrast essays, 10 descriptive essays, 10 narrative essays, and 10 opinion essays.

2.2.1. Instrument

The IELTS exam is a significant assessment of English language skills designed to provide proof that test-takers possess the linguistic abilities required by the test user for the specific language context in which they are expected to perform successfully (IELTS, 2019). Advanced to be used in this test to assess writing skills, this research utilized the "IELTS TASK 2 Writing band descriptors (public version)" (IELTS, 2023) as the scoring criteria for the essays. This tool evaluates written texts using four equally important analytical assessment criteria spread across nine performance levels, which include task response, coherence and cohesion, lexical resource, and grammatical range and accuracy. Citing Davies (2008), Pearson (2022) underscored that this feature demonstrates the way that the evaluation of IELTS writing adopts a theoretical approach centered around general proficiency, applicable to various academic domains. Furthermore, Mizumoto and Eguchi (2023) outlined that the rubric enables a comprehensive evaluation covering factors such as the handling of the task, coherence and cohesion, lexical proficiency, and grammatical variety and precision, all assessed on a 10-point scale (band) from 0 to 9. The researchers also noted that this rubric was more favorable over the other renowned rubrics, such as the 5-point scale found in the TOEFL iBT test independent writing rubrics, because the 10-point scale allows for a more detailed evaluation and facilitates a more refined distinction among scores. Taking these points into consideration, this rubric was found to be useful for the purposes of this study and it was chosen as the essay scoring instrument.

2.2.2. Human assessment

A group of three experienced EFL instructors were recruited to assess the collected essays based on the rubric selected. These instructors, experienced in teaching writing, had been working at a university-level preparatory program. One of them had five years of teaching writing experience, one of them had four years, and the other one had three years of experience in EFL contexts. The instructors followed the standardized assessment criteria of “IELTS TASK 2 Writing band descriptors (public version)” to score the essays. Three instructors rated each essay separately for the sake of inter-rater reliability. After the instructors scored the essays, the three scores generated for each essay were examined to determine if there was a score gap more than 20% (1.8 points) of the total score an essay could be assigned. For instance, if an essay was assigned 6, 8, and 9 by the instructors, the instructors were asked to re-evaluate that specific essay to ensure the inter-rater agreement and increase the reliability of the human scores. Out of the 50 essays, only 4 essays required this practice. After this process, the mean scores of the three human scores were calculated for each essay for further analysis.

2.2.3. ChatGPT assessment

A recent version of ChatGPT, ChatGPT-4o mini, was utilized to assess the same set of essays. After several releases of the model such as ChatGPT-3 and ChatGPT-3.5, an enhanced version, ChatGPT-4 is currently in public use, and it is fully available through a subscription service. However, a scaled-down version, ChatGPT-4o mini, is accessible for free to users who sign up for an account. Many people benefit from it because it is freely accessible, which is the reason why this version was selected to be under scrutiny in this study.

On a similar basis with human raters, the essays were scored by the ChatGPT-4o mini system based on the same rubric of “IELTS TASK 2 Writing band descriptors (public version)” for three times with a different prompt for each. Adapting the methodology of Mizumoto and Eguchi (2023, p. 5), the following prompt was typed in ChatGPT’s chat box to generate the first set of scores:

> I would like you to mark a/an [type of the essay] essay written by a B2-level of English as a foreign language learner. The prompt given to the learners for this essay task was [the prompt of the essay task]. The essay should be assigned a rating of 0 to 9, with 9 being the highest and 0 the lowest. You don’t have to explain why you assign that specific score. Just report a score only. The essay is scored based on the following rubric.
>
> [IELTS rubric in a plain text format.]
>
> ESSAY:
>
> [The essay in a plain text format.]

Following the generation of the first set of scores, paraphrased versions of the first prompt were used to re-evaluate the essays. The second prompt was as follows:

> I would like you to grade a/an [type of the essay] which was written by an English as a foreign language learner at B2 level. The instruction given to the learner was [the prompt of the essay task]. This essay should be given a score between 0 to 9 (0 as the lowest and the 9 as the highest). I would like only the score, not the reason why you assigned that score. Rate the essay based on the following rubric.
>
> [IELTS rubric in a plain text format.]
>
> ESSAY:
>
> [The essay in a plain text format.]

Finally, the third set of scores was generated by the subjection of the following prompt:

> I would like you to rate a/an [type of the essay] written by a B2-level English as a foreign language learner. The prompt given was: [the prompt of the essay task]. Assign a score between 0 and 9 (0 being the lowest, 9 being the highest). Only provide the score without an explanation. Use the following rubric to evaluate the essay.
>
> [IELTS rubric in a plain text format.]
>
> ESSAY:
>
> [The essay in a plain text format.]

These three different but very similar prompts were used for each essay to be assigned a score. After three scores were generated for each essay through these prompts, the scores were examined to determine if there was a considerable gap between the three scores (a score difference more than 20% or 1.8 points), a procedure which was also conducted with the human raters to increase the inter-rater agreement and the reliability of the scores. Eventually, no such instances were found. In fact, ChatGPT-4o mini assigned relatively consistent scores across the three assessments. Specifically, it assigned the same score for 31 essays in all three of the assessments. For the other 19 essays, the score differences were minor in that a re-evaluation was not regarded. Finally, the mean scores of the three ChatGPT scores were calculated for comparative analysis.

2.3. Data Analysis

The scores given by human raters and the ChatGPT system were compared to determine the degree of agreement or difference between the two assessment methods. The same set of essays was evaluated by both human raters and ChatGPT, therefore, the data were paired. In paired data, each data point in one group (human scores) corresponds directly to a data point in the other group (ChatGPT scores) because both sets of scores are for the same set of essays. In line with this, the Wilcoxon signed-rank test was operated to determine if there was a statistically significant difference between the two assessment outcomes. Willard (2020) pointed out that the Wilcoxon signed-rank test serves as an appropriate nonparametric alternative to the related samples t-test when parametric assumptions are not met, and it can be applied to matched pairs design. In the case of this study, the Wilcoxon signed-rank test was appropriate because of two main reasons: (1) the data consisted of matched pairs (each essay was assigned a score generated by both human raters and ChatGPT) and (2) the parametric assumptions of normality and variance for a paired samples t-test were not satisfied.

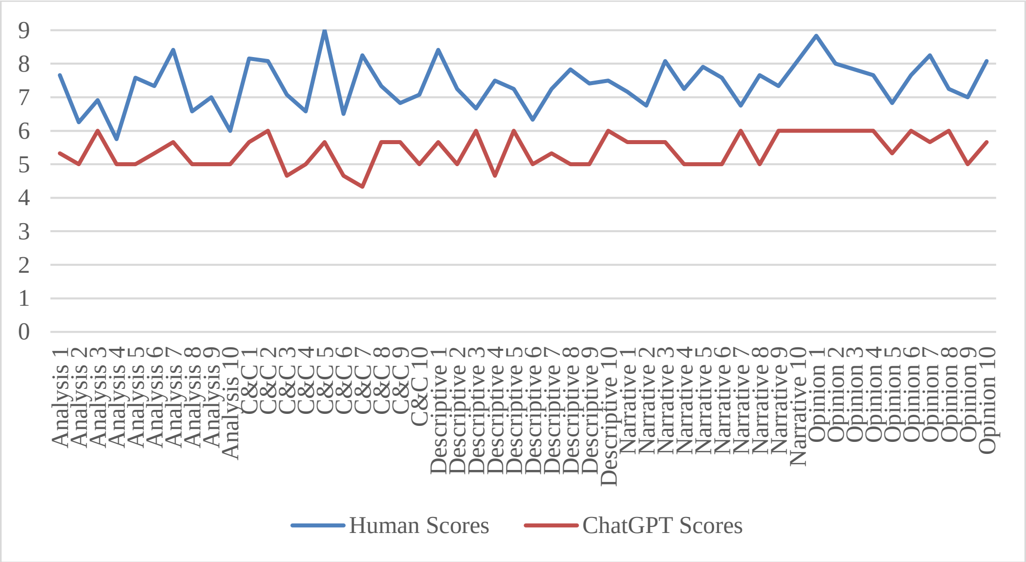
In addition to the Wilcoxon signed-rank test, the Spearman correlation coefficient was computed to reveal the direction of the relationship between the two groups of scores. Correlation tests are operated to reveal the level of the relationship between the data sets (Larson-Hall, 2012). The Spearman correlation coefficient, commonly referred to as Spearman’s rho, serves as the nonparametric alternative to the Pearson correlation coefficient (Willard, 2020). Because the current set of data violated the parametric test assumptions, the

Spearman correlation coefficient test was operated instead of the Pearson correlation test. Finally, the results of these statistical tests were presented and interpreted accordingly.

3. FINDINGS

The first research question aimed to determine if there is a significant difference between the scores generated by human raters and ChatGPT. Results of the Wilcoxon signed-rank test indicated that there is a significant difference between the human scores (Median = 7.3, N = 50) and the ChatGPT scores (Median = 5.5, N = 50), $Z = -6.1504$, $p < 0.001$, with a large effect size ($r = -0.8698$). This set of data indicates that human scores are significantly higher than those assigned by ChatGPT. This finding can be visually represented in Figure 1, where the scores generated by the human raters and ChatGPT for each essay are presented.

Figure 1. *The scores assigned by human raters and ChatGPT for each essay.*



The calculated Z value was -6.1504, and the corresponding p-value was less than 0.001. This result is statistically significant at the $p < 0.05$ level, suggesting that there is a statistically significant difference between the scores assigned by human raters and those assigned by ChatGPT. These findings indicate that the evaluations provided by humans significantly differ from those given by ChatGPT, with humans scoring higher on average. This highlights the potential discrepancy between human and ChatGPT scoring in essay evaluations. The effect size value (r) in the context of the Wilcoxon signed-rank test measures the strength of the relationship between the two samples being compared. In the realm of the current data, the large negative effect size ($r = -0.8698$) suggests that the difference in scores between the human raters and ChatGPT is not only statistically significant but also practically meaningful. This indicates a substantial gap in grading performance between the two scoring parties.

Additionally, the Spearman correlation coefficient test was operated to assess the direction of the relationship between the scores. The results yielded a Spearman correlation coefficient r_s = 0.30493, with a p-value of 0.0313. Respectively, the value of $r_s = 0.30493$ indicates a moderate positive correlation between the scores assigned by human raters and ChatGPT in a significant way ($p < 0.05$). This suggests that as the scores from human raters increase, the scores from ChatGPT tend to increase as well. While this is the case in that human and ChatGPT scores tend to rise together, the correlation is not strong enough to imply that they are interchangeable or that they assess the essays in the same manner. The correlation indicates that there is a tendency for higher human scores to align with higher ChatGPT scores, but it also reflects the variability in the scores assigned by both evaluators. Overall, it can be concluded that the findings suggest a significant difference in the evaluation of essays between human raters and ChatGPT.

The second research question aimed to reveal if the genre of the essay plays a significant role in the agreement between the scores produced by human raters and ChatGPT. To investigate whether the genre of the essay influences the agreement and correlation between the scores assigned by human raters and ChatGPT, the same set of tests was conducted for five distinct essay genres each of which contained ten essays: analysis essay, compare and contrast essay (the abbreviation C&C is used in Table 1), descriptive essay, narrative essay, and opinion essay. The results of the Wilcoxon signed-rank tests and the Spearman correlation coefficients for each genre are presented in Table 1.

Table 1. *Comparison of the human and ChatGPT scores based on the genre of the essay.*

	Wilcoxon signed-rank test					Spearman correlation	
	Human *(Median)*	ChatGPT *(Median)*	*Z*-value	*p*-value	Effect size (*r*)	r_s value	*p*-value
Analysis Essay	7	5	-2.7539	0.005	-0.8709	0.56891	0.08611
C&C Essay	7.2	5.3	-2.7539	0.005	-0.8709	0.27777	0.43713
Descriptive Essay	7.3	5.2	-3.0973	0.001	-0.9794	-0.12224	0.73656
Narrative Essay	7.5	5.7	-3.0973	0.001	-0.9794	-0.15504	0.66888
Opinion Essay	7.7	6	-2.7557	0.005	-0.8714	0.32313	0.36244

The Wilcoxon signed-rank test for the analysis essays revealed a significant difference between human and ChatGPT scores, with a *Z*-value of -2.7539 and a *p*-value of 0.005, indicating a statistically significant difference. The median human score (Median = 7) was higher than the median score assigned by ChatGPT (Median = 5), showing that human raters consistently rated the essays higher than ChatGPT. The effect size (r = -0.8709) suggests a strong difference between the two scoring systems. The Spearman correlation coefficient (r_s = 0.56891) indicates a moderate positive relationship between human and ChatGPT scores, but this relationship is not statistically significant (p = 0.08611). Although the scores tend to move in the same direction, this correlation is not strong enough to indicate substantial agreement between human and ChatGPT scores. In other words, while there is some alignment in the scoring, human raters still rated the analysis essays significantly higher than ChatGPT.

For the compare and contrast essays, the Wilcoxon signed-rank test again showed a significant difference between human and ChatGPT scores. The *Z*-value of -2.7539 and a *p*-value of 0.005 indicate a statistically significant disagreement between the two sets of scores. Human scores had a higher median (Median = 7.2) compared to ChatGPT (Median = 5.3), and the large effect size (r = -0.8709) suggests a strong disparity in how human raters and ChatGPT evaluated the essays. Furthermore, the Spearman correlation coefficient (r_s = 0.27777) indicates only a weak positive correlation between human and ChatGPT scores, and this correlation is not statistically significant (p = 0.43713). This result suggests that there is minimal alignment between human and ChatGPT scores for the compare and contrast essays.

Similarly, the test results for the descriptive essays revealed a significant difference between the human and ChatGPT scores, with a *Z*-value of -3.0973 and a *p*-value of 0.001. The median human score was 7.3, while the median ChatGPT score was considerably lower at 5.2. The effect size (r = -0.9794) is very large, indicating a very strong difference between the scores given by human raters and those assigned by ChatGPT. This suggests that ChatGPT systematically scored descriptive essays much lower than human raters did. The Spearman correlation coefficient ($r_s = -0.12224$) shows a very weak negative correlation between human and ChatGPT scores, and this relationship is not statistically significant (p = 0.73656). This result indicates almost no agreement between the scores given by human raters and those assigned by ChatGPT.

For narrative essays, the results again showed a significant difference between human and ChatGPT scores, with a Z-value of -3.0973 and a p-value of 0.001. The median human score was 7.5, while ChatGPT's median score was 5.7. The effect size (r = -0.9794) indicates a very large difference between the two scoring methods, suggesting that human raters consistently rated narrative essays higher than ChatGPT. The Spearman correlation coefficient (r_s = −0.15504) also indicates a weak negative correlation between human and ChatGPT scores, with no statistically significant relationship (p = 0.66888). This lack of significant correlation suggests that there is no meaningful alignment between human and ChatGPT scores in narrative essays. The results suggest that while human raters scored these essays higher, ChatGPT struggled to evaluate them in a similar way, further highlighting the differences in how the two systems interpret this genre.

The test results for opinion essays also revealed a significant difference between human and ChatGPT scores, with a Z-value of -2.7557 and a p-value of 0.005. The median human score was 7.7, while the median score assigned by ChatGPT was 6.0, indicating that human raters consistently assigned higher scores. The effect size (r = -0.8714) was large, showing a strong disagreement between human and ChatGPT scores. The Spearman correlation coefficient (r_s = 0.32313) revealed a weak positive correlation between human and ChatGPT scores, but this relationship was not statistically significant (p = 0.36244). This suggests that while there is a slight tendency for human and ChatGPT scores to increase together, the correlation is too weak to assert a meaningful agreement.

Overall, the results indicate that there is a statistically significant difference between human and ChatGPT scores across all essay types, with human raters consistently assigning higher scores than ChatGPT. The large effect sizes observed in all genres suggest that these differences are notable. ChatGPT appears to particularly struggle with more subjective essay types, such as descriptive and narrative essays, where the gap between human and ChatGPT scores is widest. On the other hand, the Spearman correlation results show that there is little to no meaningful correlation between human and ChatGPT scores across most essay types. The moderate positive correlation in the analysis essay suggests some degree of alignment, but the relationship is not statistically significant. For more subjective genres like descriptive and narrative essays, there is no agreement between the two scoring methods, with weak or even negative correlations observed. Eventually, it can be asserted that the genre of the essays does not play a significant role in the agreement between the scores generated by human raters and ChatGPT.

4. DISCUSSION

The findings from this study revealed significant insights into the capabilities and limitations of ChatGPT, or specifically, ChatGPT-4o mini, as an AES tool for assessing essays written by EFL learners. Building on the current findings, while ChatGPT exhibits some degree of alignment with human evaluators, substantial differences in scoring persist, echoing the concerns highlighted in previous research.

The first research question addressed the differences between the scores assigned by ChatGPT and human raters. The results indicated a statistically significant difference, with human raters consistently awarding higher scores. This finding aligns with Huang (2014), who noted that AES systems often fail to capture the nuanced elements of writing that human raters prioritize, such as creativity, depth of analysis, and linguistic fluency. Human raters' higher evaluations may suggest that they appreciate qualitative aspects that may elude AI algorithms, which typically emphasize mechanical correctness over stylistic manners.

The observed difference can also be attributed to the proficiency level of the essays analyzed. Given that this study focused on essays from B2-level learners, it is probable that it involved a level of complexity that AES tools, including ChatGPT, are unable to assess. This is consistent with Wang and Bai (2021), who found that AES systems often underperform with higher-quality writing. These findings suggest a fundamental limitation of ChatGPT: while it may be adept at detecting surface-level errors, it struggles to evaluate the intricate features of more sophisticated writing, which was also evident in earlier AES tools.

Despite the significant differences in the scores, a moderate positive correlation was found between the human and ChatGPT scores. This suggests that while ChatGPT does not replicate human evaluations precisely, it can recognize some general trends in writing quality. This finding resonates with Yancey *et al*. (2023), who reported that ChatGPT-4 could closely match human scoring under certain conditions, indicating that while AI tools may not replace human judgment, they can serve as useful adjuncts in assessing writing quality. However, as this study has shown, the current capabilities of ChatGPT-4o mini are relatively limited.

The second research question examined whether the genre of the essay influenced the agreement between the human and ChatGPT scores. The results indicated consistent differences across all five genres, supporting the assertion that genre does not mitigate the gap between human and ChatGPT evaluations. However, the results showed that the correlation between human and ChatGPT scores varied depending on the essay genre. Notably, this variation in score discrepancies by genre suggests that certain types of essays pose more significant challenges for ChatGPT assessment. The descriptive essays, which garnered the strongest difference, highlight that ChatGPT has difficulty in evaluating writing that relies on sensory details and evocative language. This finding is partly in line with Guo and Wang (2024), who noted that while ChatGPT provided detailed feedback, it occasionally delivered off-topic suggestions due to its specific prompting. Conversely, the strongest correlation was observed in analysis essays, where structured reasoning is more readily assessed by ChatGPT, suggesting that it performs better in genres where explicit criteria are more easily defined.

Overall, the findings from this study underscore that while ChatGPT shows promise as an AES tool, its limitations render it unsuitable as a standalone grading system for EFL essays. The significant differences between human and ChatGPT scores, along with the variable performance across different genres, indicate that ChatGPT is not yet capable of providing evaluations that align closely with human scoring. This can be supported by the conclusions of a very recent study by Bui and Barrot (2024), who highlighted the need for further refinement of AI algorithms and a greater sensitivity to qualitative writing aspects. Currently, while ChatGPT holds the potential to contribute to essay scoring, its limitations necessitate a cautious approach to its implementation in educational contexts. It should primarily be viewed as a supplementary tool rather than a replacement for human raters.

5. CONCLUSION

In this study, it was aimed to explore the feasibility and effectiveness of using ChatGPT, a LLM developed by OpenAI, as an AES tool for evaluating EFL learners' essays. The investigation involved comparing the scores generated by human raters to those assigned by the ChatGPT-4o mini version across various essay types. The findings revealed a significant difference between human and ChatGPT scores, with human raters assigning consistently higher scores. While a small positive correlation was observed, indicating a tendency for scores to increase together, the weak relationship suggested limited agreement between the two assessment methods. Furthermore, it was found that the genre of the essays was not a parameter mitigating between human and ChatGPT scores. While ChatGPT may show promise as an AES tool, its limitations are evident. The observed disparities highlight the complexities of language and subjective interpretation in EFL essays, which pose challenges for current AI models.

5.1. Limitations

The current study was conducted with the account of 50 essays in total, which may be argued to be a small sample. Additionally, the study included a set of essays which were written by

B2-level learners. It is suggested for further research to enrich the sample along with different proficiency levels and to include research questions accordingly.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number**: Sivas Cumhuriyet University, Educational Sciences Ethics Committee, 24.05.2024-431192.

Contribution of Authors

All stages of the study were conducted with equal contribution from both authors.

Orcid

Ahmet Can Uyar https://orcid.org/0000-0003-2438-9877
Dilek Büyükahıska https://orcid.org/0000-0002-4370-7626

REFERENCES

Almusharraf, N., & Alotaibi, H. (2022). An error-analysis study from an EFL writing context: Human and automated essay scoring approaches. *Technology, Knowledge and Learning, 28*, 1015-1031. https://doi.org/10.1007/s10758-022-09592-z

Attali, Y. (2013). Validity and reliability of automated essay scoring. In M.D. Shermis & J.C. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 181-198). Routledge.

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.* https://doi.org/10.1007/s10639-024-12891-w

Chen, H., & Pan, J. (2022). Computer or human: a comparative study of automated evaluation scoring and instructors' feedback on Chinese college students' English writing. *Asian-Pacific Journal of Second and Foreign Language Education, 7*(34), 1-20. https://doi.org/10.1186/s40862-022-00171-4

Coghlan, D., & Brydon-Miller, M. (2014). *The SAGE encyclopedia of action research*. SAGE.

Creswell, J.W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches.* SAGE Publications.

Davies A. (2008). Assessing academic English language proficiency: 40+ years of U.K. language tests. In Fox J., Wesche M., Bayliss D., Cheng L., Turner C.E., Doe C. (Eds.), *Language testing reconsidered* (pp. 73–86). University of Ottawa Press.

Guo, K., & Wang, D. (2024). To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. *Education and Information Technologies, 29*, 8435–8463. https://doi.org/10.1007/s10639-023-12146-0

Huang, S.J. (2014). Automated versus human scoring: A case study in an EFL context. *Electronic Journal of Foreign Language Teaching, 11*(1), 149-164.

IELTS. (2019). Guide for educational institutions, governments, professional bodies and commercial organisations. Cambridge Assessment English, The British Council, IDP Australia. https://www.ielts.org/-/media/publications/guide-for-institutions/ielts-guide-for-institutions-2015-uk.ashx

IELTS. (2023). IELTS Task 2 Writing band descriptors (Public version). https://takeielts.britishcouncil.org/sites/default/files/ielts_writing_band_descriptors.pdf

Larson-Hall, J. (2012). How to run statistical analyses. In A. Mackey & S.M. Gass (Eds.), *Research methods in second language acquisition: A practical guide* (pp. 245-274). Wiley-Blackwell.

Manap, M.R., Ramli, N.F., & Kassim, A.A.M. (2019). Web 2.0 automated essay scoring application and human ESL essay assessment: A comparison study. *European Journal of English Language Teaching, 5*(1), 146-162. https://doi.org/10.5281/zenodo.3461784

Mason, O., & Grove-Stephenson, I. (2002). Automated free text marking with paperless school. In M. Danson (Ed.), *Proceedings of the Sixth International Computer Assisted Assessment Conference* (pp. 216–222). Loughborough: Loughborough University.

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics, 2*, 1-13. https://doi.org/10.1016/j.rmal.2023.100050

Page, E. (1966). The imminence of ... grading essays by computer. *Phi Delta Kappan, 47*(5), 238–243.

Parker, J.L., Becker, K., & Carroca, C. (2023). ChatGPT for automated writing evaluation in scholarly writing instruction. *Journal of Nursing Education, 62*(12), 721-727. https://doi.org/10.3928/01484834-20231006-02

Pearson, W.S. (2022). Student Engagement with Teacher Written Feedback on Rehearsal Essays Undertaken in Preparation for IELTS. *Sage Open, 12*(1). https://doi.org/10.1177/21582440221079842

Wang, J., & Bai, L. (2021). Unveiling the scoring validity of two Chinese automated writing evaluation systems: A quantitative study. *International Journal of English Linguistics, 11*(2), 68-84. https://doi.org/10.5539/0jel.v11n2p68

Willard, C.A. (2020). *Statistical methods: An introduction to basic statistical concepts and analysis*. Routledge.

Yancey, K.P., Laflair, G., Verardi, A., & Burstein, J. (2023). Rating short L2 essays on the CEFR scale with GPT-4. In E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, & T. Zesch (Eds.), *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 576-584). Retrieved October 2, 2024, from https://aclanthology.org/2023.bea-1.49

Zribi, R., & Smaoui, C. (2021). Automated versus human essay scoring: A comparative study. *International Journal of Information Technology and Language Studies, 5*(1), 62-71.