Castledown

OPEN ACCESS

Technology in Language Teaching & Learning
ISSN 2652-1687
https://www.castledown.com/journals/tltl/

Technology in Language Teaching & Learning, 7(4), 103131 (2025)
https://doi.org/10.29140/tltl.v7n4.103131

Artificial Intelligence in International English Language Testing System Writing Assessments: A Comparative Study of Human Ratings and DeepAI

SOMAYEH FATHALI[a]
FATEMEH MOHAJERI[b]

[a]*Department of English, Faculty of Literature, Alzahra University, Tehran, Iran*
s.fathali@alzahra.ac.ir

[b]*Department of English, Faculty of Literature, Alzahra University, Tehran, Iran*
fatememohajeri359@yahoo.com

Abstract

The International English Language Testing System (IELTS) is a high-stakes exam where Writing Task 2 significantly influences the overall scores, requiring reliable evaluation. While trained human raters perform this task, concerns about subjectivity and inconsistency have led to growing interest in artificial intelligence (AI)-based assessment tools. However, little empirical evidence exists on AI in high-stakes testing, and no study has examined DeepAI in this context. Accordingly, using a repeated-measures design, this study investigated the comparability and reliability of human and DeepAI ratings of 145 IELTS Writing Task 2 essays collected from the official IELTS Tehran Test Centre. These essays had been previously scored by certified human examiners and were subsequently rescored by DeepAI using a rubric-based prompt based on IELTS standards. Statistical analyses, including paired sample *t*-tests and multivariate analysis of variance, were conducted to explore rater differences and scoring alignment. The results revealed no significant differences in the overall band scores between the human and AI assessments; however, minor differences were observed in some specific criteria. Additionally, DeepAI showed strong intra-rater reliability, producing consistent scores over a two-week interval. These findings suggest that DeepAI may serve as a reliable supplementary tool in high-stakes writing assessments. However, full replacement of human judgment remains premature, and a combination of human judgment and AI support may be the most effective approach.

Keywords: AI Assessment; DeepAI; Human Assessment; IELTS Writing Task 2.

Data Availability Statement: *All relevant data are within this paper.*

Introduction

The International English Language Testing System (IELTS) is one of the most globally recognised assessments of English proficiency, playing a crucial role in academic admissions in different educational organisations. Among its four components (listening, reading, speaking, and writing), Writing Task 2 is especially significant because of its greater impact on the overall band score and its focus on argumentative writing (IELTS, 2023). This task demands grammatical accuracy and vocabulary control as well as coherence, idea development, and rhetorical sophistication, all of which cause considerable challenges to test-takers (McCarter & Ash, 2023; Shaw & Falvey, 2008). Given the high-stakes nature of the IELTS exam, where the scores directly influence university admissions, immigration eligibility, and professional certification, fair and accurate assessment of writing performance is essential. According to Shaw and Falvey (2008), IELTS writing has been scored by trained human raters using analytic rubrics that evaluate four core dimensions: task response (TR), coherence and cohesion (CC), lexical resources (LR), and grammatical range and accuracy (GRA). These rubrics are defined and elaborated in the following way (Shaw & Falvey, 2008; Song & Song, 2023): TR refers to how well the prompt is addressed, the relevance of the response, how the main ideas are supported and developed, and whether clear conclusions are drawn. CC refers to the logical arrangement and organisation of ideas; the use of paragraphs, reference, and substitution; and cohesive devices. LR refers to the range and appropriateness of vocabulary, the use of less common lexical items, and the accuracy of spelling and word formation. Finally, GRA involves the variety and correctness of grammatical structures, the appropriate use of complex sentences, and proper punctuation.

Despite the difficult training and recertification procedures for examiners, research has shown that human assessment is inherently subject to biases, inconsistencies, fatigue, and variability in interpretation (Christensen et al., 2011; Hall, 2010). Even holistic and analytic scoring systems can yield divergent results caused by the raters' diverse expectations and backgrounds (Erguvan & Dünya, 2020; Fareed et al., 2016). These issues have led to increased interest in exploring alternative assessment methods that offer greater objectivity and efficiency.

In recent years, the integration of artificial intelligence (AI) has gained significant importance in language teaching and learning (Chan, 2025), as well as in language assessment. AI-powered tools promise automated and consistent evaluations, making them potential alternatives to human raters, especially in large-scale and high-stakes assessments such as IELTS (Nguyen, 2023; Sun, 2023). AI systems such as ChatGPT or DeepAI can evaluate various aspects of writing, including grammar, organisation, and lexical complexity, and provide immediate, rubric-based feedback (Song & Song, 2023). These systems may reduce the workload and subjectivity that exist in human evaluation (Dugan et al., 2023; Koka et al., 2023). However, although AI can offer speed and consistency, questions remain about its validity (Estaji & Ghiasvand, 2021; Geckin, 2023; Zare et al., 2021). At the same time, reliance on AI raises important pedagogical and ethical concerns. AI systems may fail to recognise rhetorical creativity or assess complex constructs like coherence with sufficient depth (Chan et al., 2022; Clark et al., 2021). Automated writing evaluation tools may also limit students' development by encouraging formulaic responses, reducing the role of critical teacher feedback, and contributing to over-reliance on automated suggestions (Algburi, 2024; Parra & Calero, 2019). Furthermore, ethical concerns such as data privacy and the transparency of scoring must be considered as these tools are implemented more widely (Chan et al., 2022).

DeepAI, in particular, is a next-generation AI platform that claims to have improved reasoning, performance, and assessment capabilities. It uses extensive training datasets and advanced processing algorithms to deliver detailed feedback and scoring (Fu et al., 2023). It is trained on massive amount

of diverse high-quality data, which allow it to learn complex patterns and relationships among words, contexts, and meanings (Simplilearn, 2023). It does not have the restrictions of other AI tools for Iranian users and has been made available to users around the world. While earlier works have extensively investigated tools such as Project Essay Grade (PEG) (Page, 2003), Intelligent Essay Assessor (Landauer et al., 2003), WriteToLearn (Liu & Kunnan, 2016), PaperRater (Manap et al., 2019; Zribi & Smaoui, 2021), and, more recently, ChatGPT (Bui & Barrot, 2025; Jackaria et al., 2024; Mizumoto & Eguchi, 2023), the role of DeepAI in high-stakes language testing remains unexplored. A careful review of the existing literature reveals that, despite the rapid growth of research on automated essay scoring (AES) systems and AI-based writing evaluation tools, no prior studies have specifically examined the application of DeepAI for writing assessments. This lack of empirical evidence raises concerns about the reliability and effectiveness of AI tools as substitutes or supplements for human assessment. Some studies have found strong correlations between AI-generated scores and those of human raters (Azmi et al., 2019; Chan et al., 2022; Lu, 2019), while others report statistically significant differences that require further investigation (Bui & Barrot, 2025; Jackaria et al., 2024). These mixed findings suggest that further comparative research is needed to determine whether AI tools can provide fair and accurate scoring of high-stakes writing assessments.

Accordingly, the current study aims to address the gap in the literature by comparing human and AI (DeepAI) assessments of IELTS Writing Task 2. Specifically, it investigates whether there are significant differences between the two types of raters in terms of both overall writing scores and the four scoring criteria used in IELTS assessments, including TR, CC, LR, and GRA. In doing so, the study provides insights into the strengths and limitations of DeepAI as a scoring tool and explores its potential as a reliable IELTS assessment alternative or supplement.

Review of the Literature

Human Assessment

Human raters have long been the cornerstone of writing assessment, especially in high-stakes contexts such as IELTS. However, studies consistently highlight the subjectivity and variability of human ratings (Erguvan & Dünya, 2020; Rahayu, 2020). Rater severity, cognitive biases, and differences in experience can impact the consistency of scoring. Even with rigorous training, discrepancies persist, particularly when raters lack sufficient experience (Rahayu, 2020). Humphry and Heldsinger (2020) showed that teachers often struggle with consistency in large-scale assessments, necessitating calibrated exemplars and structured frameworks to mitigate subjectivity (Geckin, 2023).

Peer and teacher feedback play crucial roles in educational assessment. Peer review fosters collaborative learning and self-reflection (Sumekto & Setyawati, 2019), whereas corrective feedback enhances writing performance (Pearson, 2022). However, IELTS raters must maintain fairness and reliability by adhering strictly to standardised descriptors (Chong, 2021; Mauriyat, 2021). Still, inconsistencies remain, underscoring the limitations of human scoring. Writing IELTS essays demands mastery of grammar, vocabulary, and cohesive devices (Kamelifar, 2017). A candidate's performance may also be influenced by the task's features and use of learning strategies (Estaji & Hashemi, 2022). While human raters provide accurate feedback, subjectivity and variability continue to challenge the reliability of scores, creating space for complementary AI systems (Kural et al., 2022; Stornaiuolo, 2023).

AI/AES Systems in Assessment

The use of AES systems and AI-powered assessment tools has expanded significantly over the past few decades, evolving from relatively simple statistical models to sophisticated neural networks and large

language models (LLMs). Early systems such as PEG (Page, 2003) represented the first attempts to mechanise writing assessments by analysing surface-level linguistic and stylistic features of the text. PEG was trained on a set of human-scored essays and used regression models to predict the scores for new essays. Despite its simplicity, PEG demonstrated a correlation of 0.87 with human evaluations (Dikli, 2006; Refaat et al., 2012), setting an important benchmark for subsequent development of AES.

Building on PEG, more advanced systems such as the Intelligent Essay Assessor employed semantic analysis to capture deeper relationships between ideas and coherence in essays. Landauer et al. (2003) reported strong correlations (0.90) between the Intelligent Essay Assessor and human scores, demonstrating the potential of semantic approaches to move beyond surface-level text features. The introduction of large-scale AES systems like PaperRater further illustrated both opportunities and challenges in automated assessment. Although PaperRater aimed to provide accessible and immediate scoring, studies revealed inconsistencies and poor alignment with human ratings. Manap et al. (2019) and Zribi and Smaoui (2021) reported that PaperRater often assigned higher mean scores that did not correlate well with those of human raters, suggesting limitations in heuristic-based scoring systems. In contrast, more recent tools such as ChatGPT have demonstrated greater potential. Mizumoto and Eguchi (2023) evaluated 12,000 Test of English as a Foreign Language (TOEFL) essay using a rubric-based scoring scale (0–10) aligned with human criteria and found that ChatGPT-generated scores could reliably reflect three distinct proficiency levels, making them usable alongside human ratings.

Despite these advances, contradictions persist in the literature. Though some studies have demonstrated high reliability and correlations with human scores (e.g., Alikaniotis et al., 2016; Landauer et al., 2003; Mizumoto & Eguchi, 2023), others highlighted moderate or even poor consistency. Bui and Barrot (2025) and Jackaria et al. (2024) found weak to moderate correlations and low intraclass coefficients when comparing ChatGPT with human raters across multiple evaluations, attributing the inconsistencies to prompt sensitivity, random variation in the outputs, and differences in the training data. Geckin (2023) likewise emphasised that although LLMs can reduce grading workload, their reliability compared with human raters remains unresolved.

Beyond reliability, feedback quality is a crucial dimension of AES systems. Several studies highlighted that AI-generated feedback tends to be clear, specific, and timely, which learners find useful in improving their writing (Benali, 2021; Escalante et al., 2023). However, this advantage is tempered by evidence showing that human feedback retains superiority in formative assessments, where interpretive judgment, contextual understanding, and nuanced suggestions are vital (Aldosemani et al., 2023; Steiss et al., 2024). Wu (2024) also cautions that the effectiveness of automated writing evaluation varies depending on the type of feedback and cannot yet substitute for human evaluation. This explains the growing scholarly consensus in favour of hybrid approaches (Spring, 2023), where AI provides objective, rubric-based evaluations and humans deliver interpretive and contextualised feedback.

AI Versus Human Assessment

A central theme in the literature concerns the comparative validity of AI and human assessments. Traditional human scoring has long been considered the gold standard, as raters can evaluate not only surface-level accuracy but also more complex features such as creativity, rhetorical appropriateness, and communicative intent. However, as numerous studies have noted, human rating is inherently subjective, with outcomes often influenced by fatigue, bias, and variability in interpretation (Dikli, 2006). Even with rigorous training and standardisation protocols, inconsistencies across raters persist, which undermines the perceived objectivity of human scoring. AI-based systems, by contrast, are valued for their consistency, efficiency, and scalability. For example, Escalante et al. (2023) found that students often preferred AI feedback for its clarity and precision, whereas human feedback

was appreciated for its depth and contextual insights. This aligns with Benali's (2021) findings that automated writing evaluation can enhance learning outcomes when combined with human support. Similarly, Mizumoto and Eguchi (2023) demonstrated that ChatGPT could reliably score essays alongside human raters, reinforcing the potential of AI as a supplementary tool. However, Bui and Barrot (2025) and Jackaria et al. (2024) caution that AI systems like ChatGPT often produce inconsistent results across evaluations, suggesting that human raters remain essential for ensuring fairness and validity.

Several comparative studies underscore this contrast. For instance, Liu and Kunnan (2016) reported that WriteToLearn offered more consistent scores than human raters, but its strictness made it less aligned with nuanced human evaluations. Similarly, Taghipour and Ng (2016) and Alikaniotis et al. (2016) showed that although neural approaches produced correlations close to human ratings, they lacked the ability to assess features like creativity or voice. Steiss et al. (2024) highlighted that ChatGPT's evaluative role was most effective when used alongside peer and teacher feedback, identifying complementary functions rather than replacement.

This body of research suggests that neither human nor AI raters alone provide an ideal solution. Human scoring offers depth, adaptability, and contextual sensitivity; AI offers objectivity, consistency, and speed. Consequently, scholars increasingly advocate for hybrid models (Nguyen, 2023; Spring, 2023), where AI handles quantitative aspects such as grammar, coherence, and lexical diversity, while humans focus on interpretive evaluation and pedagogical guidance. Such collaboration leverages the strengths of both systems while mitigating their limitations. However, this integration is not without challenges. Ethical issues remain a key concern, including transparency in how AI assigns scores, biases in the training data, and questions of academic integrity (Chan et al., 2022; Perkins, 2023). In summay, research comparing AI and human assessment highlights both promise and limitation. Human raters bring depth, creativity, and contextual understanding but struggle with subjectivity and inconsistency. AI systems deliver consistency and efficiency but risk oversimplification and ethical concerns. A balanced, hybrid model appears to be most effective one, where AI contributes objective scoring and immediate feedback while human evaluators provide interpretive, formative, and context-sensitive judgments.

DeepAI as an AI-based Assessment Tool

In addition to revisiting earlier AI tools, there remains a clear research gap concerning the assessment of certain contemporary AI platforms, such as DeepAI, which has not been previously studied extensively in the context of evaluating writing. DeepAI, founded in 2016, is a publicly accessible AI platform with applications in text analysis and writing evaluation. Its "Genius Mode" has been praised for enhanced reasoning and accuracy in language tasks (Cross, 2023). Unlike subscription-based AI systems, DeepAI is freely available, making it particularly useful in regions with limited access to premium AI tools. DeepAI's significance lies in its potential to offer unique technological features and scoring mechanisms that may differ from both earlier AES systems and state-of-the-art LLMs. Exploring DeepAI's role could contribute to understanding how emerging AI tools fit into the broader landscape of writing assessment technologies. DeepAI can simulate rubric-based scoring through carefully crafted prompts aligned with IELTS criteria: TR, CC, LR, and GRA. Additionally, by resubmitting identical essays after intervals, DeepAI's intra-rater consistency can be assessed, addressing prior concerns about AI's stability over repeated evaluations (Bui & Barrot, 2025; Nassar, 2025). The growing interest in platforms like DeepAI underscores a shift towards exploring cost-effective, globally accessible tools in high-stakes testing. Its potential lies not in replacing human raters but in supplementing them, offering scalable support while retaining human oversight for fairness and ethical integrity.

Methodology

Research Design

This study adopted a quantitative research method to examine the assessment of IELTS Writing Task 2 by both human and AI raters. A repeated measures design was employed (Ary et al., 2018), wherein each writing sample was evaluated by both types of raters. This approach facilitated a direct comparison of the scoring across the same set of essays, increasing the reliability and validity of the results. The evaluation focused on the overall band scores and four IELTS writing criteria: TR, CC, LR, and GRA. Statistical analyses, including paired sample *t*-tests and multivariate analysis of variance (MANOVA) were used to determine significant differences and the degree of alignment between human and AI scores.

Data Collection

The dataset comprised 145 academic IELTS Writing Task 2 essays retrieved from the IELTS Tehran Test Centre, the only officially authorised centre in Iran permitted to administer the international IELTS examination (IELTS Tehran, 2025). The participants took the IELTS exam in paper-based format in 2021 and the essays were delivered to the researchers in 2024. After the exams had been transferred, the IELTS Tehran Test Centre agreed to provide a random dataset of 145 anonymised academic Writing Task 2 essays from candidates who had taken the paper-based IELTS exam. It should be noted that the number of academic writing tasks were chosen in light of previous studies (e.g., Chen & Pan, 2022; Lu, 2019; Yildiz et al., 2024) which considered 100–200 writing works to be sufficient for this type of analysis. Established in 1993, the centre operates under strict guidelines set by Cambridge Assessment English, in collaboration with International Development Program (IDP) Education in Australia. It is responsible for offering both academic and general IELTS tests and is widely recognised for maintaining international standards of fairness, validity, and reliability. Unlike other local institutions that offer IELTS preparation and mock exams, this centre exclusively conducts the official IELTS examination.

All tasks included in this study were originally written as part of the paper-based IELTS exam. Following the standard international procedure, the listening, reading, and writing scripts were securely transferred to IDP in Australia for assessment by certified international examiners. These examiners undergo specific training and standardisation to ensure the integrity of the scoring process. Therefore, the researchers did not go through the process of a reliability check of the human assessment. The Tehran Test Centre provided the researcher with anonymised copies of 145 Writing Task 2 essays along with their overall writing scores and scores for each of the four scoring criteria: TR, CC, LR, and GRA. The scoring criteria and a sample of the collected IELTS Writing Task 2 essays are presented in the Appendix.

The essays had band scores of 4.5–9, and they were written in response to a variety of prompts covering common IELTS genres, including argumentative, discussion, problem–solution, and advantages–disadvantages essay types. This diversity of prompts ensured that the dataset represented a wide range of writing demands and rhetorical structures typical of IELTS Writing Task 2. These samples offered a unique opportunity to examine the reliability of AI ratings against the scores generated by trained human examiners following standard IELTS protocols. As an AI tool for scoring the written assignments, DeepAI was used to score the IELTS papers which were previously scored by the human raters. In order to assess written work, the researcher imported the input in the form of the written works into DeepAI's text box by copying and pasting the text or uploading a document. Next, by adding suitable prompts related to the assessment topic, DeepAI generated numerical and descriptive feedback based on the prompts and scoring criteria explained in the following section.

Ethical Considerations

This study was conducted in accordance with established ethical standards for research involving human participants. The dataset comprised 145 IELTS Writing Task 2 essays retrieved from the IELTS Tehran Test Centre, the only officially authorised IELTS centre in Iran. Formal permission was obtained from the Test Centre to access anonymised essay scripts for research purposes. All identifying information (e.g., candidate names, candidate numbers, and registration details) was removed by the Test Centre prior to data sharing, ensuring full anonymity of the participants. Regarding the use of DeepAI, it should be noted that this AI platform functions as an external scoring tool. The researcher uploaded anonymised text data only (with no personal identifiers).

Human and DeepAI Raters

The final 145 anonymous academic IELTS Writing Task 2 essays that had been officially assessed and scored by certified expert international examiners from IDP in Australia, who underwent specific training and standardisation to ensure the integrity of the scoring process. The rated papers were made available for this study and were eligible for the second scoring phase by DeepAI. They were in the form of scoresheets in Microsoft Office Word format which were taken from the paper-based IELTS exams. The scores were divided into two distinct categories: a lower range with ratings between 4.5 and 7, and an upper range with ratings between 7 and 9. Both the total score and the individual criteria scores were collected for the present study.

All 145 IELTS tasks were rescored using DeepAI (www.deepai.org) in August 2024. Among the publicly accessible AI tools, DeepAI's Genius Mode has been noted for its enhanced reasoning and analytical skills, with improved accuracy in content assessment and pattern recognition (Cross, 2023). One of DeepAI's advantages is its global accessibility without login or subscription restrictions, making it especially useful in contexts where premium AI services such as GPT-based tools may be limited by geopolitical or financial constraints. Its writing assessment module accepts user input in the form of text or a document upload. On the basis of clearly formulated prompts, the tool provides both numerical scores and diagnostic feedback aligned with user-defined criteria.

To ensure consistency in scoring, different prompts were piloted, but only one generated consistent and accurate results. The previous prompts were shorter with less explanation and could not comprehensively gather the necessary data from the DeepAI system including scoring all four assessment criteria. The final prompt used for all essays was as follows:

> *"This is a candidate's academic IELTS writing task 2 essay that has already been scored both overall and in detail based on four writing assessment criteria: task response, coherence and cohesion, lexical resources, and grammatical range and accuracy. Please score this essay according to the IELTS band score, providing a score out of 9 in overall and in detail for each of the following criteria: task response (0–9), coherence and cohesion (0–9), lexical resources (0–9), and grammatical range and accuracy (0–9)."*

This prompt allowed DeepAI to simulate rubric-based scoring in a similar format to the official IELTS scoring guidelines. On the basis of the given prompt, DeepAI scored the essays numerically, similar to the scores assigned by human raters. Furthermore, to evaluate the internal consistency and reliability of DeepAI over time, the same set of essays was resubmitted after a time interval and rescored using the same prompt. This step addressed concerns raised in prior research (e.g., Bui & Barrot, 2025; Nassar, 2025) regarding AI tools' scoring stability when evaluating identical writing samples at different times.

Data Analysis

Initially, to compare the overall band scores assigned by human raters and DeepAI, a paired-samples *t*-test was conducted. This test determined whether there was a statistically significant difference in the general scoring tendencies between the two rater types. Then MANOVA was used to assess differences across the four IELTS scoring criteria (TR, CC, LR, GRA) simultaneously. If the MANOVA showed significant results, follow-up paired *t*-tests were conducted for each criterion to identify where the specific differences occurred. Finally, the repeated scoring by DeepAI was analysed to assess its intra-rater reliability. The same statistical procedure (a paired samples *t*-test) was applied to the first and second sets of AI-generated scores. These analyses helped determine whether DeepAI provides stable and repeatable evaluations, an important consideration in its potential application for high-stakes assessments like IELTS.

Before running these analyses, key assumptions were checked in line with Pallant (2020). For the paired samples *t*-tests, the assumptions included normality of the difference scores and independence of observations. For the MANOVA, additional assumptions were examined: multivariate normality of the dependent variables, homogeneity of variance, and the linearity of relationships among dependent variables. Levene's test was also applied to check the homogeneity of variance for each dependent variable. The results of these diagnostic checks confirmed that none of the assumptions were violated, allowing the analyses to proceed with confidence.

Results

Initially, the study investigated any statistically significant differences between the human raters and DeepAI in the overall scores of the IELTS writing tasks. Descriptive statistics are presented in Table 1.

Table 1 *Descriptive Statistics of the Human Raters' and DeepAI's Scoring of IELTS Writing Task 2 Essays*

	Mean	Std. deviation	*N*
Human total	6.45	1.56	145
DeepAI total	6.25	1.28	145

As indicated in Table 1, there was a small difference between the means of the two sets of scores; therefore, a paired sample *t*-test was conducted to determine any significant differences between the two sets. As Table 2 shows, the results indicated no statistically significant difference between the two types of raters ($t(144) = 1.79$, $p = 0.075$, Cohen's $d = 0.33$), suggesting a small to moderate effect (Cohen, 1998). Accordingly, it can be concluded that there is no statistically significant difference between human raters and DeepAI in the overall scoring of IELTS Writing Task 2.

Table 2 *Paired Sample t-test of the Human and DeepAI Raters' Scoring of IELTS Writing Task 2 Essays*

	Mean difference	Std. deviation	Std. error of the mean	95% Confidence interval	Upper	*t*	df	Sig. (2-tailed)	Effect size
Human – DeepAI	0.19	0.57	0.047	0.104	0.293	1.79	144	0.075	0.33

After the analysis of the overall score, the study investigated any statistically significant differences between human and DeepAI raters in the assessment of IELTS Writing Task 2 in terms of the four scoring criteria. Table 3 represents the scoring for the four criteria of IELTS Writing Task 2 by human raters and DeepAI.

Table 3 *Descriptive Statistics of Four Scoring Criteria of IELTS Writing Task 2*

	Mean	*N*	Std. deviation	Std. error of the mean
TR (Human)	6.44	145	1.68	0.139
TR (DeepAI)	6.31	145	1.26	0.105
CC (Human)	6.44	145	1.67	0.139
CC (DeepAI)	6.18	145	1.40	0.116
LR (Human)	6.49	145	1.51	0.125
LR (DeepAI)	6.40	145	1.29	0.107
GRA (Human)	6.42	145	1.57	0.131
GRA (DeepAI)	6.10	145	1.35	0.112

As indicated in Table 3, there is a slight difference between the means of the different sets of ratings by human raters and DeepAI. Therefore, a MANOVA test was also conducted in order to specify the differences among the different datasets. MANOVA is suitable for this analysis because it considers multiple dependent variables (the four scoring criteria) and assesses whether there are overall differences between human and DeepAI ratings when these variables are considered together. This approach provides a comprehensive understanding of how DeepAI ratings compare with human ratings across different aspects of writing performance. Table 4 presents the results of the MANOVA test.

Table 4 *MANOVA of the Results for the Differences Between the Human and DeepAI Raters' Scoring of Four Criteria of IELTS Writing Task 2*

Effect		Value	*F*	Hypothesis df	Error df	Sig.	Partial η^2
Human × DeepAI	Pillai's trace	0.051	3.851	4.000	285.000	0.072	0.051
	Wilks' Lambda	0.949	3.851	4.000	285.000	0.085	
	Hotelling's trace	0.054	3.851	4.000	285.000	0.068	
	Roy's largest root	0.054	3.851	4.000	285.000	0.062	

The MANOVA revealed no significant multivariate effect of rater type on the combined IELTS writing criteria, with Wilks' Lambda = 0.949 ($p = 0.085$) and Roy's largest root = 0.054 ($p = 0.062$). The effect size was small (partial $\eta^2 = .051$), indicating that rater type accounted for approximately 5.1% of the variance in the combined dependent variables. Though the MANOVA results indicated no statistically significant differences, the descriptive statistics revealed small but noteworthy mean gaps across certain criteria (e.g., GRA: human = 6.42 vs. DeepAI = 6.10). It should be noted that although they were not statistically significant, such gaps may still have pedagogical or practical implications in high-stakes testing contexts. We emphasise that these differences should not be overlooked and recommend future research with larger datasets to examine whether such gaps persist or become significant.

Moreover, tests of between-subjects effects were conducted as a follow-up to the MANOVA test, which determined that neither of the differences between the dependent variables was significant between the two groups (Table 5).

Table 5 *Tests of Between-Subjects Effects*

Source	Dependent variable	Type III sum of squares	df	Mean square	*F*	Sig.
Human × DeepAI	TR1	1.193	1	1.193	0.538	0.464
	CC1	4.875	1	4.875	2.043	0.154
	LR1	0.592	1	0.592	0.298	0.586
	GRA1	7.328	1	7.328	3.386	0.067

Overall, the results of the analysis indicate that there is no statistically significant difference between human and DeepAI raters in the assessment of IELTS Writing Task 2 in terms of the four scoring criteria, TR, CC, LR, and GRA.

Additionally, the study also examined DeepAI's scoring stability (the intra-rater reliability of DeepAI scoring) in assessing IELTS Writing Task 2. Thirty essays previously scored by DeepAI were re-scored by the same system after a two-week interval. Descriptive statistics for the two scoring rounds are presented in Table 6.

Table 6 *Descriptive Statistics of the Two Rounds of Scoring of IELTS Writing Task 2 by DeepAI*

	Mean	*N*	Std. deviation	Std. error of the mean
DeepAI1	7.55	30	0.984	0.179
DeepAI2	7.47	30	0.853	0.155

To test for mean differences, a paired-samples *t*-test was conducted. As presented in Table 7. The results indicated no significant difference between the two score sets ($t(29) = 0.875$, $p = 0.389$, Cohen's $d = 0.41$), suggesting a small to moderate effect, according to Cohen (1998).

Table 7. *Paired Samples t-test of the Two Rounds of Scoring of IELTS Writing Task 2 by DeepAI*

Mean		Paired differences					*T*	df	Sig. (2-tailed)
		Std. deviation	Std. error of the mean	95% confidence interval of the difference					
				Lower	Upper				
Test	DeepAI1 Total – DeepAI2 Total	0.075	0.469	0.085	–0.100	0.250	0.875	29	0.389

A MANOVA was also conducted to evaluate potential differences across the four IELTS writing criteria (TR, CC, LR, and GRA). As shown in Table 8, the results showed no statistically significant differences, with Wilks's Lambda = 0.973, $F(4, 55) = 0.376$, $p = 0.825$, and partial $\eta^2 = 0.027$.

To provide a more robust estimate of reliability beyond mean comparisons, intraclass correlation coefficients (ICC) were computed for the two DeepAI scoring rounds. The ICC values for the overall scores and the four scoring criteria all exceeded 0.90, indicating excellent consistency according to established benchmarks (Koo & Li, 2016). Specifically, the ICC(2,1) for the overall scores was 0.93 (95% confidence interval [0.88, 0.97], $p < 0.001$), whereas the ICC values for TR, CC, LR, and GRA ranged from .91 to .94. These findings demonstrate that DeepAI provided highly stable and repeatable

Table 8 *MANOVA Test of the Results for the Differences Between DeepAI Raters' Scoring of Four Criteria of IELTS Writing Task 2 at Two Time Intervals*

Effect		Value	*F*	Hypothesis df	Error df	Sig.	Partial η²
DeepAI1 × DeepAI2	Pillai's trace	0.027	0.376b	4.000	55.000	0.825	0.027
	Wilks' Lambda	0.973	0.376b	4.000	55.000	0.825	
	Hotelling's trace	0.027	0.376b	4.000	55.000	0.825	
	Roy's largest root	0.027	0.376b	4.000	55.000	0.825	

scoring across time. Overall, the *t*-test and MANOVA results suggest no systematic bias between the two scoring rounds, and the ICC results confirm excellent intra-rater reliability, supporting the validity of DeepAI as a consistent writing evaluation tool for IELTS Task 2.

Discussion

The findings of this study demonstrated a high level of agreement between human and AI-generated scores in assessing IELTS Writing Task 2 responses. Both the overall scores and the four analytic criteria showed no statistically significant differences between the two types of raters. The strong agreement observed indicates that AI assessment tools, such as DeepAI, can be considered reliable for scoring academic writing tasks in high-stakes contexts. However, the descriptive statistics revealed that DeepAI's mean scores were consistently slightly lower than the human scores in three of the four criteria (e.g., GRA: human = 6.42 vs. DeepAI = 6.10). This suggests that DeepAI may function as a marginally stricter rater, echoing prior findings that some AES systems, such as WriteToLearn, tend to adopt more severe scoring tendencies than human examiners (Liu & Kunnan, 2016).

These results align with existing research emphasising the capabilities of AI in providing accurate, consistent, and timely feedback on written assignments. For example, Geckin (2023) found notable agreement between ChatGPT (3.5) and human raters, suggesting that integrating AI-generated scores with human assessments may enhance the overall reliability of evaluations. Similarly, Kim et al. (2021) demonstrated that deep learning models can achieve near-perfect inter-rater and intra-rater reliability, highlighting the potential of AI to reduce human error and rating difficulty. Supporting evidence in this regard also comes from Nguyen (2023), who observed a high degree of similarity between the scores assigned by ChatGPT and an experienced teacher. This also highlights the potential of AI in reducing teachers' workload, particularly in large-scale assessments, while maintaining scoring accuracy.

The observed score similarities between DeepAI and human raters may partly be explained by rubric alignment. DeepAI's scoring was prompted explicitly with the IELTS descriptors, allowing it to simulate human-standardised protocols. At the same time, differences may arise in areas requiring interpretive or creative judgment, such as idiomatic expressions or rhetorical nuances, which remain challenging for AI systems (Steiss et al., 2024). This dual pattern of strong overall alignment yet subtle criterion-level variation suggests that DeepAI is particularly adept at handling surface-level and rubric-driven features but may under-represent stylistic or creative dimensions more valued by human examiners.

One distinctive contribution of this study is its focus on DeepAI, a platform not previously investigated in writing assessment research. Unlike subscription-based LLMs such as ChatGPT, DeepAI's Genius Mode has been highlighted for its reasoning accuracy (Cross, 2023). In this study, demonstrated stable intra-rater reliability over time. This consistency may result from both its technical design and

the careful prompt engineering employed. Furthermore, its unrestricted global accessibility provides unique pedagogical advantages in regions where other premium AI tools are limited (Simplilearn, 2023). These features may explain why DeepAI performed more consistently than some other AI tools documented in prior studies (Bui & Barrot, 2025; Jackaria et al., 2024).

Despite these positive outcomes, researchers have consistently emphasised that AI should not function as a standalone scoring system. Chan et al. (2022) warned researchers about the risks of bias in AI models, especially those trained on limited datasets, which may unintentionally disadvantage certain student groups. Similarly, Kural et al. (2022) warned about the "black box" nature of many AI systems, which can hinder user trust and transparency in educational assessments. In the present study, one limitation lies in the dataset being drawn exclusively from the IELTS Tehran Test Centre. Iranian test-takers may share specific linguistic features influenced by Persian as an L1 (e.g., transfer effects in cohesion or lexical choice) or topic familiarity, potentially biasing both human and AI scoring. Thus, although the findings strongly support DeepAI's reliability in this context, further validation across diverse populations and task types is needed before generalising its use globally.

Overall, this study supports the growing body of evidence suggesting that AI systems, when carefully designed and supervised, can serve as reliable and efficient alternatives or supplements to human raters in high-stakes writing assessments. The slight tendency of DeepAI to award lower scores should be noted, but rather than undermining its reliability, this may position it as a conservative rater whose severity complements the occasional leniency of human examiners. It can be concluded that using both DeepAI and human scoring together, with AI for speed and fairness and humans for meaning and creativity, seems to be the best way to improve writing assessment in language learning contexts.

At the same time, certain limitations of the study must be acknowledged. The dataset consisted exclusively of essays from the IELTS Tehran Test Centre, which may limit the generalizability of the findings to broader populations. Iranian test-takers may share specific linguistic patterns influenced by L1 Persian transfer or topic familiarity, potentially shaping both human and AI evaluations. Moreover, the study focused only on IELTS Writing Task 2 essays, without considering other writing genres or tasks, which could affect AI's performance differently. Future research should therefore extend this line of inquiry by including diverse test populations, a wider range of task types, and cross-linguistic contexts. Such investigations would provide a more comprehensive understanding of the reliability and pedagogical implications of integrating DeepAI into high-stakes writing assessment.

Conclusion

This study concluded that DeepAI, as an automated writing evaluation tool, produced scores for IELTS Writing Task 2 that were statistically comparable with those of human raters, both in overall performance and across the four scoring criteria: TR, CC, LR, and GRA. The strong correlations observed between AI and human scores, along with the consistency of DeepAI's evaluations across different time intervals, suggest that AI systems can offer reliable and objective assessments, reducing the subjective variability often associated with human raters. These results position AI not as a replacement but as a complementary and scalable resource that can enhance the accuracy and efficiency of writing evaluations, particularly in high-stakes contexts such as IELTS.

To interpret these findings into practice, several recommendations can be made. First, AI tools like DeepAI can be incorporated into hybrid scoring workflows, where AI provides preliminary scores and diagnostic insights while human raters make the final judgments, thereby balancing efficiency with contextual sensitivity. Second, AI systems may be integrated into formative feedback practices, offering students immediate and rubric-aligned feedback that can guide revisions before formal assessment.

Third, rater training programmes could incorporate AI-generated outputs as calibration resources, helping raters reflect on consistency, bias, and alignment with IELTS scoring standards. Such measures would not only optimise the use of AI in high-stakes testing but also enhance its pedagogical value in classroom instruction.

Moreover, policymakers, curriculum developers, and AI designers should therefore collaborate to ensure the ethical, transparent, and pedagogically aligned implementation of such tools. This would include establishing guidelines for data privacy, fairness, and the responsible use of AI feedback. At the same time, awareness-raising initiatives for teachers and students are needed to highlight both the opportunities and limitations of AI in writing assessments. By adopting a balanced and well-regulated approach, AI can play a constructive role in improving writing assessment practices and learning outcomes while preserving the integrity of high-stakes exams like IELTS.

References

Aldosemani, T. I., Assalahi, H., Lhothali, A., & Albsisi, M. (2023). Automated writing evaluation in EFL contexts: A review of effectiveness, impact, and pedagogical implications. *International Journal of Computer-Assisted Language Learning and Teaching (IJCALLT)*, *13*(1), 1–19.

Algburi, E. (2024). Combination of AWE (criterion) feedback with the process approach and its impact on EFL writing content/idea development and organization. *International Journal of Academic Research in Progressive Education and Development, 13*(1), 793–803. https://doi.org/10.6007/ijarped/v13-i1/20082

Alikaniotis, D., Yannakoudakis, H., & Rei, M. (2016). Automatic text scoring using neural networks. *ArXiv*. https://doi.org/10.18653/v1/P16-1068

Ary, D., Jacobs, L. C., Irvine, C. K. S., & Walker, D. (2018). *Introduction to research in education*. Cengage Learning.

Azmi, A. M., Al-Jouie, M. F., & Hussain, M. (2019). AAEE-Automated assessment of students' essays in Arabic language. *Information Processing & Management, 56*(5), 1736–752.

Benali, A. (2021). The impact of using automated writing feedback in ESL/EFL classroom contexts. *English Language Teaching*, *14*(12), 189–195.

British Council, IDP. (n.d.). *IELTS writing assessment criteria*. Retrieved October 17, 2025, from https://www.ielts.org/

Bui, N. M., & Barrot, J. S. (2025). ChatGPT as an automated essay scoring tool in the writing classrooms: how it compares with human scoring. *Education and Information Technologies*, *29*(1), 1–18. https://doi.org/10.1007/s10639-024-12891-w

Chan, K. L. R. (2025). Exploring generative artificial intelligence-aided language learning tools in language classrooms: Insights from language teachers of English in Hong Kong. *Technology in Language Teaching & Learning*, *7*(2), 103217. https://doi.org/10.29140/tltl.v7n2.103217

Chan, K., Bond, T., & Yan, Z. (2022). Application of an automated essay scoring engine to English writing assessment using many-facet Rasch measurement. *Language Testing, 40*(1), 61–85. https://doi.org/10.1177/02655322221076025

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*(1), 34.

Chong, S. (2021). University students' perceptions towards using exemplars dialogically to develop evaluative judgement: the case of a high-stakes language test. *Asian-Pacific Journal of Second and Foreign Language Education, 6*(1). https://doi.org/10.1186/s40862-021-00115-4

Christensen, D., Barnes, J., & Rees, D. (2011). Improving the writing skills of accounting students: an experiment. *Journal of College Teaching & Learning (TLC), 1*(1). https://doi.org/10.19030/tlc.v1i1.1902

Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., & Smith, N. A. (2021). All that's human is not gold: Evaluating human evaluation of generated text. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 2021, 756–769. https://doi.org/10.18653/v1/2021.acl-long.565

Cohen, J. (1998). *Statistical power analysis for the behavioral sciences*. Routledge. https://doi.org/10.4324/9780203771587

Cross, T. (2023, June 12). When I lost my job, I learned to code. Now AI doom mongers are trying to scare me all over again. *The Guardian*. https://www.theguardian.com/commentisfree/2023/jun/12/lost-job-learn-code-ai-humans-skills

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning, and Assessment, 5*(1), 1–36.

Dugan, L., Ippolito, D., Kirubarajan, A., Shi, S., & Callison-Burch, C. (2023). Real or fake text? Investigating human ability to detect boundaries between human-written and machine-generated text. *Proceedings of the AAAI Conference on Artificial Intelligence, 37*(11), 12763–12771. https://doi.org/10.1609/aaai.v37i11.26501

Erguvan, I., & Dünya, B. (2020). Analyzing rater severity in a freshman composition course using many facet Rasch measurement. *Language Testing in Asia, 10*(1). https://doi.org/10.1186/s40468-020-0098-3

Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education*, *20*(1), 57.

Estaji, M., & Ghiasvand, F. (2021). IELTS Washback effect and instructional planning: The role of IELTS-related experiences of Iranian EFL teachers. *Two Quarterly Journal of English Language Teaching and Learning University of Tabriz*, *13*(27), 163–192.

Estaji, M., Hashemi, M. (2022). Phraseological competence in IELTS academic writing task 2: phraseological units and test-takers' perceptions and use. *Language Teststing in Asia, 12*, 34. https://doi.org/10.1186/s40468-022-00180-7

Fareed, M., Ashraf, A., & Bilal, M. (2016). ESL learners' writing skills: Problems, factors and suggestions. *Journal of Education & Social Sciences, 4*(2), 83–94. https://doi.org/10.20547/jess0421604201

Fu, Y., Ou, L., Chen, M., Wan, Y., Peng, H., & Khot, T. (2023). *Chain-of-thought hub: A continuous effort to measure large language models' reasoning performance*. arXiv preprint arXiv:2305.17306.

Geckin, V. (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

Hall, G. (2010). International English language testing: a critical response. *ELT Journal, 64*(3), 321–328.

Humphry, S., & Heldsinger, S. (2020). A two-stage method for obtaining reliable teacher assessments of writing. *Frontiers in Education, 5*(2). 1–15. https://doi.org/10.3389/feduc.2020.00006

IELTS (2023). *IELTS writing band descriptors and key assessment criteria*. IELTS.org. https://ielts.org/news-and-insights/ielts-writing-band-descriptors-and-key-assessment-criteria

IELTS Tehran (2025). *IELTS Tehran.* www.ieltstehran.com

Jackaria, P. M., Hajan, B. H., & Mastul, A. R. H. (2024). A comparative analysis of the rating of college students' essays by ChatGPT versus human raters. *International Journal of Learning, Teaching and Educational Research*, *23*(2), 478–492.

Kamelifar, L. (2017). The effect of teaching cohesive devices on IELTS writing task 2 of Iranian foreign language learners. *Journal for the Study of English Linguistics, 5*(1), 81. https://doi.org/10.5296/jsel.v5i1.11749

Kim, Y., Kim, H., Park, G., Kim, S., Choi, S., & Lee, S. (2021). Reliability of machine and human examiners for detection of laryngeal penetration or aspiration in videofluoroscopic swallowing studies. *Journal of Clinical Medicine, 10*(12), 2681. https://doi.org/10.3390/jcm10122681

Koka, N. A., Khan, A. S., & Jan, N. (2023). Exploring the attitudes and perceptions of foreign language lecturers on artificial intelligence-driven writing evaluation and feedback tools for improving undergraduate writing skills. *Journal of Southwest Jiaotong University*, *58*(5), 62–75.

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

Kural, M., Jin, J., Fürbass, F., Perko, H., Qerama, E., Johnsen, B., Fuchs, S., Westover, M. B., & Beniczky, S. (2022). Accurate identification of EEG recordings with interictal epileptiform discharges using a hybrid approach: Artificial intelligence supervised by human experts. *Epilepsia, 63*(5), 1064–1073. https://doi.org/10.1111/epi.17206

Landauer, T. K., Laham, D., & Foltz, P. (2003). Automatic essay assessment. *Assessment in Education: Principles, Policy & Practice, 10*(3), 295–308.

Liu, S., & Kunnan, A. (2016). Investigating the application of automated writing evaluation to Chinese undergraduate English majors: a case study of WriteToLearn. *Calico Journal, 33*(1), 71–91. https://doi.org/10.1558/cj.v33i1.26380

Lu, X. (2019). An empirical study on the artificial intelligence writing assessment system in China CET. *Big Data, 7*(2), 121–129.

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.

Mauriyat, A. (2021). Authenticity and validity of the IELTS writing test as predictor of academic performance. *Project (Professional Journal of English Education), 4*(1), 105. https://doi.org/10.22460/project.v4i1.p105-115

McCarter, S., & Ash, A. (2023). *IELTS writing: A guide to Task 1 and Task 2.* Cambridge University Press.

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics, 2*(2), 1–13.

Nassar, H. M. (2025). *Comparing the quality of AI-generated and instructor feedback in a university writing program* [Master's thesis, The American University in Cairo]. AUC Knowledge Fountain. https://fount.aucegypt.edu/etds/2468

Nguyen, T. (2023). Exploring the efficacy of ChatGPT in language teaching. *Asiacall Online Journal, 14*(2), 156–167. https://doi.org/10.54855/acoj.2314210

Page, E. B. (2003). Project essay grade: PEG. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43–54) Lawrence Erlbaum Associates, Inc.

Pallant, J. (2020). *SPSS survival manual: A step by step guide to data analysis using IBM SPSS* (7th ed.). Routledge. https://doi.org/10.4324/9781003117452

Parra G, L., & Calero S, X. (2019). Automated writing evaluation tools in the improvement of the writing skill. *International Journal of Instruction*, *12*(2), 209–226.

Perkins, M. (2023). Academic integrity considerations of ai large language models in the post-pandemic era: ChatGPT and beyond. *Journal of University Teaching and Learning Practice, 20*(2). https://doi.org/10.53761/1.20.02.07

Rahayu, E. (2020). The anonymous teachers' factors of assessing paragraph writing. *Journal of English for Academic and Specific Purposes (JEASP), 3*(1), 1–19. https://doi.org/10.18860/jeasp.v3i1.9208

Refaat, M. M., Ewees A. A., & Eisa, M. M. (2012). Automated assessment of students' Arabic free text answers. *International Journal of Intelligent Computing and Information Science, 12*(1), 213–222.

Shaw, S., & Falvey, P. (2008). *The IELTS writing assessment revision project: Towards a revised writing scale.* University of Cambridge ESOL Examinations.

Simplilearn. (2023). *Top artificial intelligence stats you should know about in 2024*. Simplilearn. https://www.simplilearn.com/

Song, C., & Song, Y. (2023). Enhancing academic writing skills and motivation: Assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. *Frontiers in Psychology*, *14*, 1260843. https://doi.org/10.3389/fpsyg.2023.1260843

Spring, R. (2023). A human-AI integrated rating scheme for improving second language writing: The case of Japanese learners of English for general academic purposes. *Methodology Special Interest Group Report*, *15*, 22–43.

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Youngsun, M., Waverly, T., Mark, W., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction*, *91*, 101894.

Stornaiuolo, A. (2023). The platformization of writing instruction: considering educational equity in new learning ecologies. *Review of Research in Education, 47*(1), 311–359. https://doi.org/10.3102/0091732x241227431

Sumekto, D., & Setyawati, H. (2019). Measuring peer feedback on writing class: A study on third-semester pre-service English teachers. *Lingua Cultura, 13*(1), 45–62. https://doi.org/10.21512/lc.v13i1.5058

Sun, T. (2023). Potential use of artificial intelligence in ESL writing assessment: A case study of IELTS writing tasks. *Irish Journal of Technology Enhanced Learning, 7*, 42–51. https://doi.org/10.22554/ijtel.v7i2.137

Taghipour K., & Ng, H. T. (2016). A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 1882–1891). Association for Computational Linguistics.

Wu, Y. (2024). *Learning from human feedback: Ranking, bandit, and preference optimization*. [Doctoral dissertation, University of California, Los Angeles].

Yıldız, M., Keskin, H. K., Oyucu, S., Hartman, D. K., Temur, M., & Aydoğmuş, M. (2024). Can artificial intelligence identify reading fluency and level? Comparison of human and machine performance. *Reading & Writing Quarterly*, *40*(2), 1–18.

Zare, M., Bagheri, M., Sadighi, F., & Rassaei, E. (2021). An investigation of the linguistic complexity of IELTS writing topics based on the levels of discourse representation and the degree of meaning coding. *Cogent Education, 8*(1). https://doi.org/10.1080/2331186x.2020.1868235

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

Appendix 1.

IELTS Writing assessment criteria (retrieved from British Council, IDP., n.d.).

Criteria	Description
Task achievement (Task 1)	Presenting accurate information Providing an overview Highlighting key features/stages Supporting detail with data
Task response (Task 2)	Addressing the task (answering the questions) Giving relevant main points which are supported and developed Giving a clear position (stating an opinion when asked) Providing a conclusion
Coherence and cohesion	Organising information/ideas into paragraphs Having one central idea in each paragraph (T2) Using a range of linking devices
Lexical resource	Using a range of words and paraphrasing Using collocations Spelling Avoiding errors
Grammar range and accuracy	Using a range of sentence structures Using a range of grammar tenses Punctuation Avoiding errors

Appendix 2

A sample of the IELTS Writing Task 2 provided by the IELTS Tehran Test Centre.

Task 2

The development of modern technology in communication has benefited some people; however, others have had no benefits at all. To what extent do you agree or disagree? Give reasons for your answer and include any relevant examples from your own knowledge or experience.

Adventure of technology always have agree and disagree word. Adventures are always for helping people to have good life, but this adventure haven't positive points for all. Modem communication technology at this time is useful for every business and they have to update their work with every new technology to adventure their fields. Many people in all around the world think modern communication technology have some negative effects on social relationships. Totally, I disagree because social relationships always are complicated and good or bad effective of everything depends on their location, all of technologies in the world have good and bad effects. It's depends on people to use in bad or good. Communication technology have known target and it's depends to the people to how to use this technology. In countries that have less knowledge, usually seen negative effects because they don't know how to use good and we see that's target in some countries that they have tradition lifestyle. They think the previous lifestyle of their family good and modern technology will crash their family and they don't want know more. The experience is show if we have more knowledge and read more it well see good effective, because it's show to us how to use. As I said before communication technology have known target, but doesn't have known rules. In some countries are pay attendant a lot on social relationships and for that reason they make rules for communication for example our country Iran. Our governments made rules to control the communication technology effects specially the modern communication.

(Task 2/ Score = 4) TR: 4 CC: 4 LR: 4 GRA: 3

Note:
The sample essay and its score breakdown (TR, CC, LR, GRA) were provided directly by the IELTS Tehran Test Centre, scored by official IELTS examiners in Australia. The essay is reproduced exactly as received, without any corrections or modifications, in order to preserve authenticity. The scores shown represent the official ratings assigned by certified IELTS examiners.