13% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography

Match Groups

45 Not Cited or Quoted 11%
Matches with neither in-text citation nor quotation marks

7 Missing Quotations 2%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

9% Internet sources
10% Publications
5% Submitted works (Student Papers)

Match Groups

- **45** Not Cited or Quoted 11%
  Matches with neither in-text citation nor quotation marks
- **7** Missing Quotations 2%
  Matches that are still very similar to source material
- **0** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

9% Internet sources
10% Publications
5% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Publication	Landry Dwiyoga Daniswara. "The Influence of AI-Generated Feedback on Student... 1%
2	Internet	journal-gehu.com 1%
3	Internet	tesol.huflit.edu.vn <1%
4	Internet	nyjxxb.net <1%
5	Student papers	University of Bristol <1%
6	Internet	opus.bsz-bw.de <1%
7	Internet	brightideas.houstontx.gov <1%
8	Student papers	Swiss School of Management <1%
9	Student papers	United International College <1%
10	Internet	seipi.org.ph <1%

11	Internet	
www.researchsquare.com		<1%
12	Publication	
"Communication and Applied Technologies", Springer Science and Business Medi...		<1%
13	Internet	
learningspot.co		<1%
14	Internet	
downloads.hindawi.com		<1%
15	Internet	
leansigmacorporation.com		<1%
16	Internet	
pt.scribd.com		<1%
17	Internet	
ray.yorksj.ac.uk		<1%
18	Internet	
www.cultureready.org		<1%
19	Student papers	
University of London External System		<1%
20	Student papers	
Georgia Institute of Technology Main Campus		<1%
21	Student papers	
University of Witwatersrand		<1%
22	Internet	
www.mdpi.com		<1%
23	Internet	
www.nature.com		<1%
24	Publication	
Smruti Bulsari, Kiran Pandya. "Quantitative Research Using R", Routledge, 2026		<1%

25 Internet
journal.ump.edu.my <1%

26 Internet
www.lib.eduhk.hk <1%

27 Publication
Li Dong. "Self-regulatory Writing Strategies and Second Language Writing Profici... <1%

28 Publication
Mackenzie L. Thomas, Seyma N. Yildirim-Erbasli, Shruthi Hariharan. "Exploring un... <1%

29 Publication
Marjan Asadi, Saman Ebadi, Ahmed Rawdhan Salman, Rana Taheri, Laleh Moham... <1%

30 Publication
Nan Yang, Ning Gao, Tengda Zhang. "Artificial Intelligence in EFL Writing Enhanc... <1%

31 Internet
pdfs.semanticscholar.org <1%

32 Internet
proyectos.inei.gob.pe <1%

33 Internet
www.eltereader.hu <1%

34 Internet
www.ijiet.org <1%

35 Internet
www.researchgate.net <1%

36 Publication
"Empowering Educators: Integrating AI Tools for Personalized Language Instructi... <1%

37 Publication
"SK298 topic 7 depression WEB153412", Open University <1%

38 Publication
Sait Gürbüz, Ning Ding, Arnold Bakker. "Research Methods for Business and Man... <1%

Journal of General Education and Humanities
Vol. 5, No. 1, February 2026, pp. 1943 – 1953, https://doi.org/10.58421/gehu.v5i1.1189
ISSN 2963-7147

The Use of Artificial Intelligence in Assessing the IELTS Academic Writing Task Essays

Aliza Kamalatuzzahroh[1], Joko Priyana[2]
[1,2]Universitas Negeri Yogyakarta, Yogyakarta, Indonesia

Article Info

Article history:

Received 2026-01-23
Revised 2026-02-21
Accepted 2026-02-28

Keywords:

Artificial Intelligence (AI)
Automated Essay Scoring
IELTS Academic Writing
Writing Assessment

ABSTRACT

This study investigates the accuracy of Artificial Intelligence (AI) in assessing the IELTS Academic Writing Task essays by comparing AI-generated and human examiner scores and feedback. Despite the increasing adoption of AI-based assessment tools, limited empirical evidence exists regarding their validity and reliability in high-stakes IELTS writing evaluation. Therefore, this study aims to determine whether significant differences exist between AI and human scoring and to examine the qualitative characteristics of the feedback provided. This research employed a mixed-method explanatory design involving ten participants who completed a computer-based IELTS prediction test. Their essays were independently evaluated by an AI scoring system and a human rater using IELTS band descriptors. Quantitative analysis using a paired-sample t-test measured differences in assigned scores, while qualitative content analysis examined patterns, depth, and focus of the feedback provided. The findings indicate a statistically significant difference between AI-generated and human-assigned scores ($p = 0.022$), with a mean difference of 0.4 points, suggesting that AI tended to assign higher scores. The feedback analysis reveals that AI primarily focuses on technical aspects such as grammar, vocabulary, and sentence structure, offering general improvement suggestions, whereas human feedback demonstrates greater depth and personalization. These results suggest that while AI enhances scoring efficiency, it cannot fully replace human evaluative judgment in complex academic writing assessment.

Corresponding Author:

Aliza Kamalatuzzahroh
Faculty of Language, Arts and Culture, Universitas Negeri Yogyakarta
Email: alizakamalatuzzahroh.2022@student.uny.ac.id

1. INTRODUCTION

Among various international English proficiency examinations, the International English Language Testing System (IELTS) has become one of the most widely recognized benchmarks for academic admission, professional certification, and migration purposes worldwide [1]. Administered by the British Council, University of Cambridge Local

Examinations Syndicate, and the Australian International Development Programme (IDP) Education, IELTS is designed to measure candidates' communicative competence across listening, reading, writing, and speaking skills [2]. In particular, the Academic Writing module is considered one of the most cognitively demanding components, as it requires analytical reporting in Task 1 and argumentative essay construction in Task 2. Writing performance is evaluated using four analytic criteria: Task Achievement, Coherence and Cohesion, Lexical Resource, and Grammatical Range and Accuracy [3]. Given the high-stakes consequences of IELTS scores for individuals' educational and professional trajectories, ensuring the validity, reliability, and fairness of writing assessment remains a critical concern in language testing research [4], [5], [6].

The increasing global demand for rapid English proficiency certification has accelerated the digital transformation of language testing, including the introduction of computer-based IELTS formats and online mock examinations. This shift has led to the growing adoption of Artificial Intelligence (AI) technologies in automated writing evaluation systems. AI-powered Automated Essay Scoring (AES) systems employ natural language processing and machine learning algorithms to generate scores and feedback within a short time frame [7]. AI tools such as ChatGPT allow users to input prompts and essays, producing band estimates alongside detailed evaluative comments. Unlike official IELTS feedback, AI-generated responses are often presented in structured formats, including tables that separate evaluation according to assessment criteria [8]. These outputs frequently provide explicit grammar corrections, lexical suggestions, coherence analysis, and even revision tips for improvement. From a formative assessment perspective, such detailed and customized feedback may offer pedagogical advantages by facilitating clearer understanding and iterative revision [9]. Nevertheless, the critical question remains whether the scoring accuracy and qualitative feedback generated by AI align with the standards applied by certified human examiners in high-stakes contexts.

Theoretically, writing assessment is grounded in construct validity and performance-based evaluation frameworks, which require scoring mechanisms to accurately represent the underlying construct of academic writing ability [10]. Previous empirical studies have demonstrated the feasibility of AI-assisted scoring in language assessment contexts, primarily focusing on quantitative score agreement between AI systems and human raters. Investigations by Jiang et al., Wale, Mizumoto, and Bui and Barrot suggest that AI can provide rapid scoring and individualized feedback with moderate-to-high levels of statistical correlation [11], [12], [13], [14]. However, comparative findings remain mixed, as some studies report strong alignment while others identify discrepancies in band estimation and evaluative interpretation [15], [16], [17]. More importantly, most prior research has concentrated predominantly on scoring accuracy without systematically examining qualitative differences in written feedback content [18], [19]. Consequently, a comprehensive comparison encompassing both quantitative score agreement and qualitative feedback analysis remains underexplored in the context of IELTS Academic Writing.

Although prior studies have explored automated scoring accuracy in general ESL/EFL contexts, a clear research gap remains in systematically comparing AI-generated IELTS Academic Writing scores and qualitative feedback with those provided by certified

https://doi.org/10.58421/gehu.v5i1.1189 1945

human examiners under IELTS-specific band criteria. Many existing investigations focus primarily on technological performance metrics or pedagogical usefulness, rather than construct-aligned validity within high-stakes standardized assessment frameworks [20], [21]. Furthermore, limited research has examined how AI-generated feedback differs qualitatively from human evaluative commentary in terms of depth, diagnostic precision, and discourse-level analysis. This absence of direct comparative validation creates uncertainty regarding the extent to which AI systems can reliably replicate human rating standards in the IELTS Academic Writing assessment. Consequently, a rigorous comparative investigation is required to determine whether AI-based scoring can function as a valid complementary or alternative assessment mechanism.

Despite growing scholarly interest in AI-based assessment, no specific empirical study has simultaneously examined (1) the statistical difference between AI-generated and human-assigned IELTS scores and (2) the qualitative similarities and differences in written feedback across assessment criteria. This gap is particularly significant given that IELTS is a high-stakes examination where reliability and validity are paramount. o address this gap, this study adopts a comparative quantitative–qualitative framework that systematically examines both statistical score agreement and diagnostic feedback alignment between AI-generated outputs and certified human examiners. By integrating inter-rater reliability analysis with discourse-level feedback comparison, the study seeks to determine whether AI-based assessment can maintain construct representation and fairness within IELTS Academic Writing evaluation.

By adopting a comparative analytical framework grounded in language assessment theory, this study proposes a structured evaluation model to assess AI reliability within IELTS writing contexts. The findings are expected to contribute theoretically to discussions on construct validity in automated assessment and practically to institutions, educators, and test-preparation providers considering AI integration. Ultimately, this research aspires to clarify whether AI can enhance assessment efficiency without compromising accuracy, equity, and pedagogical integrity in high-stakes language testing environments.

2. METHOD

Within this research, a mix of quantitative and qualitative approaches was used to provide a descriptive form containing spoken or written observed information. The quantitative approach allowed for the objective measurement of AI accuracy in assessing the IELTS Academic Writing test results. By utilizing statistical methods, the researcher can determine how closely AI assessments align with human evaluations. Meanwhile, in qualitative research, the researcher performed document studies as the aspect of data collection [22]. Following the Creswell qualitative guidelines in gathering the qualitative data, the researcher followed the five phases, which consist of deciding the participants and research sites, obtaining permission, deciding the types of information to gather, creating data collection forms, and managing the process ethically. Therefore, the qualitative approach is considered suitable for this research since it needs inclusive observation from supporting documents such as the IELTS Prediction Test results, as well as the IELTS Official Test results.

The data were gathered from 10 participants' IELTS prediction test results provided on a certain website, specifically the academic writing task 1 and 2 scores, as well as the written feedback generated by an AI tool, namely ChatGPT. The human examiner was an experienced and certified IELTS examiner who had been actively involved in the industry for over seven years. Ensuring the reliability and accuracy of the information provided in this research, the chosen examiner holds official certification and possesses extensive expertise in assessing IELTS candidates, especially in the writing section.

This research used two instruments, in which the quantitative component focuses on evaluating the accuracy of AI in scoring writing tasks by comparing the CBT Prediction test results made by human examiners. Then, the results were tested using a t-test in SPSS software; meanwhile, the qualitative component involves document studies and document comparison between the AI-generated results and human feedback. Finally, to support the research result, a comparison score with the IELTS Official Test results was performed as well. Therefore, a thorough conclusion was drawn.

The quantitative data were arranged and categorized from these hypotheses: H0: There is no significant difference between the mean scores given by AI and human assessors.; H1: There is a significant difference between the mean scores given by AI and human assessors. The research applied a Paired Samples t-test to compare the means of two related groups to see if there is a significant difference between them. In this case, the researcher wants to compare the AI assessment scores to human assessment scores for the same set of writing samples.

In addition to the quantitative analysis, the researcher also seeks to gather qualitative data to contrast and compare the written feedback generated by AI and human assessors by conducting a document analysis.

In order to achieve the credibility of the data, the writer used experts' judgment from an English Education lecturer at Yogyakarta State University to verify and cross-check the data to ensure that everyone has the same understanding of the direction the researcher planned to take. Besides that, the researcher read and reread the data thoroughly to ensure the suitability of the data for the research questions. The researcher also used the triangulation technique to boost confidence in the study's findings.

3. RESULTS AND DISCUSSION

3.1. Quantitative Results

A total of 10 participants took the IELTS Academic Writing prediction test using a computer-based platform. The essays were assessed by AI (ChatGPT) and a human examiner, following these four criteria, namely Task Achievement/Response (TA/TR), Coherence and Cohesion (CC), Lexical Resource (LR), and Grammatical Range and Accuracy (GRA). The assessment was gathered in the form of scores, calculated as the average of the four aforementioned components, and finally rounded to the nearest half or whole band.

https://doi.org/10.58421/gehu.v5i1.1189

Table 1. Summary of Overall Band Scores for Each Participant

Participant Code	AI WT 1	AI WT 2	AI Overall Band	Human WT 1	Human WT 2	Human Overall Band	Score Difference
ALC	5	5	5	5	4.5	4.5	0.5
ANZ	5	5.5	5.5	4.5	5.5	5	0.5
BY	5	6	5.5	5	6.5	6	0.5
FA	5	5.5	5.5	4	5	4.5	1.0
IM	6	6.5	6.5	6.5	6	6	0.5
KE	6	6	6	6	5.5	5.5	0.5
MAJ	6	6	6	6	6	6	0.0
MM	6	6	6	6	4.5	5	1.0
RU	6	6	6	5	5.5	5.5	0.5
ZCP	5.5	5.5	5.5	5	5.5	5.5	0.0

Note: WT 1 = Writing Task 1, WT 2 = Writing Task 2, Difference = AI Overall Band – Human Overall Band

The data shows that the scores mostly differ by 0.5, as 6 participants exhibited this difference, where AI scores were higher. Meanwhile, 2 participants received a 1.0 band difference with AI scores, assigning the higher score. Eventually, only two participants had the same scores from both the AI and human examiners. This pattern demonstrates variability between AI and human examiners, where AI has a tendency to assign more generous scores.

To ensure the data met the necessary assumptions for further testing, a normality test using the Shapiro-Wilk test was conducted. Following this, Levene's test was performed to test the homogeneity of variance, and then proceeded to a paired sample t-test to evaluate the significance of the differences between the AI and human scores.

Table 2. Normality Test (Shapiro-Wilk)

Score Type	Statistic	Df	Sig.
AI Scores	.906	10	.258
Human Scores	.878	10	.124

Table 2 explains that this test follows key points where the Null Hypothesis (H0): the data is normally distributed, and the Alternative Hypothesis (H1): the data is not normally distributed. In the decision-making, the Shapiro-Wilk test suggests that if $p>0.05$, the data follows a normal distribution. In contrast, if the $p<0.05$, it means the data does not follow a normal distribution. As shown in the table, the Sig. (p-value) for the AI Score is 0.258, and the Human Score is 0.124. Therefore, the data is normally distributed for both the AI and Human scores based on the Shapiro-Wilk test, because both p-values are greater than 0.05, and it fails to reject the null hypothesis.

Furthermore, the Levene's test was performed to assess the homogeneity of variances as the key assumption later in the t-test, to check whether the variances of the different groups (AI and human scores) are equal and comparable [23]. The hypotheses for Lavene's test are H0: the variances are equal/homogenous, H1: the variances are not equal/heterogeneous. Therefore, if the $p>0.05$, meaning it fails to reject H0, or the variances

are qual (homogeneity is met); if the p<0.05, meaning it rejects H0, or the variances are not equal (homogeneity is not met).

Table 3. Test of Homogeneity of Variances

Score Result	Lavene Statistic	p-value
Based on Mean	1.357	0.259
Based on Median	0.554	0.466
Trimmed Mean	1.285	0.272

As interpreted from the shown table, all the p-value scores are >0.05, thus it fails to reject the null hypothesis, meaning the variances between AI and human scores based on the mean, median, and trimmed mean are equal and homogenous.

The final stage of the quantitative test performed in this research is the Paired Sample T-test. This test is used to determine whether there is a statistically significant difference between the paired scores from two groups. In deciding the significance, these hypotheses are used: H0: there is no significant difference between the AI and human scores, H1: there is a significant difference between the AI and human scores.

Table 4. Paired Sample T-test

Metric	Value
Mean Difference	0.400
Std. Deviation	0.459
Std. Error Mean	0.145
95% CI (Lower-Upper)	0.071-0.729
T	2.752
Df	9
p-value (2-tailed)	0.022

Table 4 shows that Column Sig. (2-tailed) Alternatively, the p-value column scores .022 which is less than the significance level of 0.05. This means it rejects the null hypothesis (H0), and it can be concluded that there is a significant difference between the AI and human scores.

3.2. Qualitative Results

Feedback from both AI and human examiners is compared across the four assessment criteria, namely the Task Achievement/Response, Coherence & Cohesion, Lexical Resource, Grammatical Range and Accuracy. After examining the qualitative data from the feedback generated by AI and human examiners for both IELTS Writing Task 1 and 2, the researcher finally identified notable similarities and differences in their assessments.

3.2.1 Key Differences

Depth of Feedback

AI has a tendency to be more technical and general, concentrating on larger topics such as cohesive device use, lexical resources, and grammatical range. Sometimes AI

feedback is not as context-aware and does not offer as many specific examples or in-depth fixes. Human feedback provides a more thorough and nuanced analysis, emphasizing concrete examples, more lucid explanations, and recommendations that can be put into practice. The human examiner frequently offers specific suggestions for how to make the essays better and tangible examples of mistakes or confusing passages.

Task Achievement & Response

The AI mainly acknowledges completion of the task but frequently makes suggestions for clarification or additions without paying as much attention to whether the data or analysis is correct and sufficient. A human examiner is usually more critical of the argument's coherence (for Task 2) or the data's completeness and accuracy (for Task 1). In order to completely fulfil the task requirements, it highlights the necessity of thorough analysis, more precise data or examples, and more focused comparisons.

Coherence and Cohesion

The AI often concentrates more on the technical elements of coherence, like the application of cohesive devices, rather than always paying as much attention to paragraph development or overall essay organization as the human examiner does. The human examiner mainly highlights how important it is to have better paragraphing, clearer transitions, and a more cogent flow between ideas. It frequently offers more useful recommendations for organizing the essay.

Lexical Resources

The AI tends to provide specific options for better word choice and instead concentrates more on broad advice like avoiding repetition or extending vocabulary. The human examiner often draws attention to more particular instances of awkward or incorrect vocabulary and frequently offers substitute words to increase variety and precision.

Grammatical Range and Accuracy

The AI feedback mainly identifies common grammatical mistakes like verb forms and subject-verb agreement, but it does not provide as many specific examples or individualized fixes. The human examiner mainly gives concrete instances of grammatical mistakes and strange wording, along with advice on how to fix them.

3.2.2 Key Similarities

Identification of Major Issues

Similar problems with the essays, such as awkward wording, repetition, ambiguous arguments, and grammatical errors, are recognized by both AI and human examiners. They both draw attention to issues with task completion, coherence, and cohesiveness, highlighting the need for more precise and illustrative data or examples.

Suggestion for Improvement

Constructive recommendations for enhancement in areas such as coherence, lexical variety, and grammatical accuracy are offered by both AI and human feedback. Both

emphasize the significance of a clearer structure, better transitions, and more precise language, even though the human feedback is more in-depth.

Focus on Key Assessment Criteria

The four main criteria for the IELTS writing assessment are covered in both types of feedback. Although the feedback from humans provides more details, both cover similar crucial topics.

Positive Feedback

Examiners using AI and humans are alike in recognizing the essays' strong points, which include their attempts to employ intricate sentence structures, some pertinent vocabulary, and a basic essay structure.

Overall, both the AI and human examiner feedback for the IELTS writing Tasks highlight essential areas for achievement, coherence, lexical resources, and grammatical accuracy. Despite the differences, both forms of feedback align in recognizing key strengths and weaknesses by providing constructive advice to enhance the quality of the writing. Ultimately, the AI feedback acts as a broad and effective tool for identifying patterns and technical improvements, whereas the human feedback is typically more contextual and personalized, since it provides a clearer roadmap for improving particular aspects of the essays. Both feedback provides a complementary method for evaluating essays, because AI offers quick insights and humans offer more detailed guidance.

3.3. Discussion

This study highlights important insights into the evolving role of AI related to the IELTS Academic Writing Tasks assessment. Generative AI tools like ChatGPT, in particular, have emerged as a promising alternative as well as a complement to human raters. The findings of this research reinforce existing studies suggesting that while AI exhibits strong capabilities in identifying grammatical mistakes [24], it lacks in accurately evaluating higher-order writing competencies such as task response, argumentation, and contextual nuance [9], [25]. The AI systems, as they are shaped by algorithmic logic, tend to accurately operationalize language assessment through quantifiable metrics such as grammar, vocabulary, coherence devices, etc. In contrast, it struggles with interpretive judgment, an area where human interpretation remains superior.

Practically, the findings underline the limitations of the AI tool as observed on the 0.4 band points discrepancy ($p = 0.022$) between AI and human scores, where AI tended to generate higher and more tentative scores when the process is repeated. This suggests potential risks if the scores were adapted as a standalone evaluation system. The statistical difference between AI and human scores is in line with previous research [12], [14], [25], which indicated weak-to-moderate correlations between both raters. This result brings out a fundamental paradigm in which AI emphasizes syntactic correctness and structural clarity, while human examiners account for creativity, persuasiveness, and communicative intent of language use into consideration. Eventually, these divergences raise critical questions about assessment validity and whether AI scores accurately represent a test-taker's ability to

communicate effectively in real-life contexts. Furthermore, the moderate yet inconsistent reliability of ChatGPT reinforces concerns that fully relying on AI alone may jeopardize the transparency and fairness as fundamental aspects of standardized testing systems.

On the other hand, the qualitative comparison of AI and human feedback revealed significant pedagogical implications that went beyond numerical scores. AI seemed to provide technical, efficient, and standardized feedback, yet it sometimes lacked depth, personalization, and contextual awareness. Human examiner, in contrast, offered holistic, supportive, comprehensive, and learner-centered feedback. This particular approach is critical to build up test-takers' motivation and improvement [26]. Feedback on task achievement and coherence-cohesion demonstrated the human examiner's superior ability in interpreting communicative intent by offering meaningful suggestions for improvement. These findings are consistent with an increasing amount of research that shows human raters are crucial for encouraging higher-order thinking and reflective revision in writing tasks [27], [28].

This study proposes a more complex, integrated assessment framework rather than viewing AI and human scoring as mutually exclusive. AI tools can efficiently handle large volumes of essays, identify common mistakes, and provide immediate feedback, especially in preliminary diagnostic testing. However, final summative assessments, especially in higher-stakes contexts, must retain human analysis to ensure fairness and ethical responsibility. Therefore, hybrid assessment models that maintain the interpretive and ethical aspects provided by human raters while using AI's strengths in consistency and speed are better combined. Such integration is more applicable than substitution.

4. CONCLUSION

This study examined the accuracy, reliability, and qualitative characteristics of AI-based assessment in evaluating IELTS Academic Writing Task essays through a comparative analysis with certified human examiners. The findings indicate that while AI demonstrates efficiency and consistency in scoring, particularly in technical linguistic features such as grammar, vocabulary, and structural organization, discrepancies remain when assessing higher-order writing constructs, including argument development, coherence at the discourse level, and contextual appropriateness. These findings confirm that AI-based systems are capable of supporting writing evaluation processes; however, they have not yet achieved full construct equivalence with human judgment in high-stakes assessment contexts. Rather than functioning as a replacement, AI currently operates more effectively as a complementary evaluative tool.

From a theoretical perspective, this study contributes to ongoing discussions on construct validity and reliability in automated writing assessment by demonstrating that statistical score alignment alone is insufficient to establish full assessment equivalence. The inclusion of qualitative feedback comparison highlights the importance of examining diagnostic depth and interpretative nuance in addition to numerical agreement. Practically, the findings provide implications for language testing institutions, educators, and test-preparation providers by suggesting that AI may be strategically integrated for formative assessment, preliminary screening, or large-scale practice testing. However, careful

alignment with official band descriptors and fairness standards remains essential to preserve assessment integrity in high-stakes examinations.

Despite its contributions, this study is subject to several limitations. The relatively small sample size and the use of a single AI tool limit the generalizability of the findings across broader populations and alternative AI systems. Additionally, the investigation was confined to IELTS Academic Writing tasks, which may not represent other writing genres or proficiency frameworks. Future research should involve larger and more diverse participant groups, compare multiple AI platforms, and incorporate advanced reliability measures such as inter-rater agreement indices. Further studies may also explore the longitudinal effects of AI-generated feedback on writing improvement and learner autonomy. For the general public, particularly IELTS candidates and educators, this research provides evidence-based insight into the strengths and limitations of AI-assisted writing assessment, promoting more informed and responsible use of emerging technologies in language learning and evaluation contexts.

REFERENCES

[1] J. Read, "Test Review: The International English Language Testing System (IELTS)," *language testing*, vol. 39, no. 4, pp. 679–694, Oct. 2022, doi: 10.1177/02655322221086211.
[2] M. A. S. Al-Malki, "Testing the Predictive Validity of the IELTS Test on Omani English Candidates' Professional Competencies," *IJALEL*, vol. 3, no. 5, Jul. 2014, doi: 10.7575/aiac.ijalel.v.3n.5p.166.
[3] P. Peltekov, "The International English Language Testing System (IELTS): A Critical Review," *JELTL*, vol. 6, no. 2, p. 395, Aug. 2021, doi: 10.21462/jeltl.v6i2.581.
[4] S. W. Chong and X. Ye, *Developing Writing Skills for IELTS: A Research-based Approach.* Routledge, 2020.
[5] W. Pearson, "A comparative study of lexical bundles in IELTS Writing Task 1 and 2 simulation essays and tertiary academic writing," *Journal of Academic Language and Learning*, vol. 15, no. 1, pp. 27–52, 2021, [Online]. Available: https://journal.aall.org.au/index.php/jall/article/download/717/435435511
[6] V. A. Veerappan and T. Sulaiman, "A Review on IELTS Writing Test, Its Test Results and Inter Rater Reliability," *TPLS*, vol. 2, no. 1, pp. 138–143, Jan. 2012, doi: 10.4304/tpls.2.1.138-143.
[7] M. Y. M. Amin, "AI and Chat GPT in Language Teaching: Enhancing EFL Classroom Support and Transforming Assessment Techniques," *Intern. j., high. educ. pedag.*, vol. 4, no. 4, pp. 1–15, Dec. 2023, doi: 10.33422/ijhep.v4i4.554.
[8] S. Fathali and F. Mohajeri, "Artificial intelligence in international English language testing system writing assessments: A comparative study of human ratings and DeepAI," *TLTL*, vol. 7, no. 4, p. 103131, Nov. 2025, doi: 10.29140/tltl.v7n4.103131.
[9] N. R. Taşkin BediZel, "Evolving landscape of artificial intelligence (AI) and assessment in education: A bibliometric analysis," *International Journal of Assessment Tools in Education*, vol. 10, no. Special Issue, pp. 208–223, Dec. 2023, doi: 10.21449/ijate.1369290.
[10] Y.-J. Lee, R. O. Davis, and S. O. Lee, "University students' perceptions of artificial intelligence-based tools for English writing courses," *ONLINE J COMMUN MEDIA TECHNOL*, vol. 14, no. 1, p. e202412, Feb. 2024, doi: 10.30935/ojcmt/14195.
[11] Z. Jiang, Z. Xu, Z. Pan, J. He, and K. Xie, "Exploring the Role of Artificial Intelligence in Facilitating Assessment of Writing Performance in Second Language Learning," *Languages*, vol. 8, no. 4, p. 247, Oct. 2023, doi: 10.3390/languages8040247.
[12] B. D. Wale, "Artificial intelligence in education: Effects of using integrative automated writing evaluation programs on honing academic writing instruction," *CP*, vol. 43, no. 1, pp. 273–287, Feb. 2024, doi: 10.21831/cp.v43i1.67715.
[13] A. Mizumoto and M. Eguchi, "Exploring the potential of using an AI language model for automated essay scoring," *Research Methods in Applied Linguistics*, vol. 2, no. 2, p. 100050, Aug. 2023, doi: 10.1016/j.rmal.2023.100050.
[14] N. M. Bui and J. S. Barrot, "ChatGPT as an automated essay scoring tool in the writing classrooms: how it compares with human scoring," *Educ Inf Technol*, vol. 30, no. 2, pp. 2041–2058, Feb. 2025, doi: 10.1007/s10639-024-12891-w.

[15] A. Alshehri, "AI's effectiveness in language testing and feedback provision," *Social Sciences & Humanities Open*, vol. 12, p. 101892, 2025, doi: 10.1016/j.ssaho.2025.101892.
[16] D. Y. H. Lee, A. J. Parker, C. F. Norbury, and D. R. Shanks, "Validating AI-assisted evaluation of open science practices in brain sciences: ChatGPT, Claude and human expert comparisons," *Royal Society Open Science*, vol. 13, no. 2, p. 250381, Feb. 2026, doi: 10.1098/rsos.250381.
[17] A. Beikian, "Evaluating AI-Driven Feedback in IELTS Writing: A Comparative Analysis of Grok and Qualified Human Examiners.," *Iranian Journal of English for Academic Purposes*, vol. 14, no. 2, 2025, doi: https://dor.isc.ac/dor/20.1001.1.24763187.2025.14.2.7.1.
[18] S. Fathali and F. Mohajeri, "Artificial intelligence in international English language testing system writing assessments: A comparative study of human ratings and DeepAI," *TLTL*, vol. 7, no. 4, p. 103131, Nov. 2025, doi: 10.29140/tltl.v7n4.103131.
[19] A. N. Sari, "Exploring the Potential of Using AI Language Models in Democratising Global Language Test Preparation," *ijte*, vol. 4, no. 4, pp. 111–126, Nov. 2024, doi: 10.54855/ijte.24447.
[20] Y. Anistyasari, S. C. Hidayati, S. Suparji, E. Ekohariadi, and D. A. Kusumaningtyas, "Comparing AI and Human Assessment of Academic Writing Skills: A Kappa Analysis," *E3S Web Conf.*, vol. 645, p. 06014, 2025, doi: 10.1051/e3sconf/202564506014.
[21] G. P. Georgiou, "Differentiating Between Human-Written and AI-Generated Texts Using Automatically Extracted Linguistic Features," *information*, vol. 16, no. 11, p. 979, Nov. 2025, doi: 10.3390/info16110979.
[22] J. W. Creswell, *Educational research: Planning, conducting, and evaluating quantitative and qualitative research (4th ed.)*. Pearson, 2012.
[23] J. L. Gastwirth, Y. R. Gel, and W. Miao, "The Impact of Levene's Test of Equality of Variances on Statistical Theory and Practice," *Statist. Sci.*, vol. 24, no. 3, Aug. 2009, doi: 10.1214/09-STS301.
[24] T. Sun, "Potential Use of Artificial Intelligence in ESL Writing Assessment: A Case Study of IELTS Writing Tasks," *telji*, vol. 7, no. 2, pp. 42–51, Dec. 2023, doi: 10.22554/ijtel.v7i2.137.
[25] R. Shabara, K. ElEbyary, D. Boraie, and TIRF (The International Research Foundation for English Language Education), "Teachers Or Chatgpt: The Issue Of Accuracy And Consistency In L2 Assessment," *TEwT*, vol. 2024, no. 2, 2024, doi: 10.56297/vaca6841/LRDX3699/XSEZ5215.
[26] R. Schmidt-Fajlik, "ChatGPT as a Grammar Checker for Japanese English Language Learners: A Comparison with Grammarly and ProWritingAid," *acoj*, vol. 14, no. 1, pp. 105–119, Jun. 2023, doi: 10.54855/acoj.231417.
[27] A. Pfau, C. Polio, and Y. Xu, "Exploring the potential of ChatGPT in assessing L2 writing accuracy for research purposes," *Research Methods in Applied Linguistics*, vol. 2, no. 3, p. 100083, Dec. 2023, doi: 10.1016/j.rmal.2023.100083.
[28] D. T. Dien, H. B. Nhu, and B. P. Thao, "Applying Chatgpt To Optimize Efl Teaching And Assessment," *EJEL*, vol. 10, no. 1, Jun. 2025, doi: 10.46827/ejel.v10i1.6084.