Teaching English as a Second Language Quarterly (TESLQ)
(Formerly Journal of Teaching Language Skills)
44(1), Winter 2025, pp. 97-117
https://doi.org/10.22099/tesl.2025.50852.3313

Online ISSN: 2981-1546
Research Paper

A Comparative Evaluation of Artificial Intelligence Scoring Versus Human Scoring of EFL Students' Essays

Vahid Reza Mirzaeian [1]

Abstract

The evaluation of students' writings and the allocation of scores are traditionally time-intensive and inherently subjective, often resulting in inconsistencies among human raters. Automated essay scoring systems were introduced to address these issues; however, their development has historically been resource-intensive, restricting their application to standardized tests such as TOEFL and IELTS. Consequently, these systems were not readily accessible to educators and learners. Recent advancements in Artificial Intelligence (AI) have expanded the potential of automated scoring systems, enabling them to analyze written texts and assign scores with increased efficiency and versatility. This study aimed to compare the efficacy of an AI-based scoring system, DeepAI, with human evaluators. A quantitative approach, grounded in Corder's (1974) Error Analysis framework, was used to analyze approximately 200 essays written by Persian-speaking EFL learners. Paired sample t-tests and Pearson correlation coefficients were employed to assess the congruence between errors identified and scores assigned by the two methods. The findings revealed a moderate correlation between human and AI scores, with AI diagnosing a greater number of errors than human raters. These results underscore the potential of AI in augmenting writing assessment practices while highlighting its pedagogical implications for language instructors and learners, particularly in evaluating the essays of EFL students.

Keywords: Artificial Intelligence, Writing, Essay, Scoring, Writing errors

Since Artificial Intelligence (AI) continues to evolve at breakneck speed, educators are tasked with the ongoing challenge of devising effective strategies to evaluate students' written assignments using this technology. Meanwhile, students themselves are in dire need of personalized, remote assistance to hone their skills in revising their own written production (Nova, 2018). In the vast array of tools, resources, and programs available to guide English as a Foreign Language (EFL) students in advancing written skills, AI systems have recently entered the scene and are poised to offer solutions to several of the obstacles and constraints associated with human essay evaluation. Utilizing AI, these cutting-edge tools are outfitted with the capability to swiftly and efficiently assess an individual's grammar and writing style while

* Received: 13/08/2024 Accepted: 01/01/2025

[1] Associate Professor, Department of English, Faculty of Literature, Alzahra University, Tehran, Iran; mirzaeian@alzahra.ac.ir

How to cite this article:
Mirzaeian, V. R. (2025). A Comparative Evaluation of Artificial Intelligence Scoring Versus Human Scoring of EFL Students' Essays. *Teaching English as a Second Language Quarterly (Formerly Journal of Teaching Language Skills), 44*(1), 97-117. doi: 10.22099/tesl.2025.50852.3313

providing valuable feedback and guidance to foster further improvement (Prinsloo & Bothma, 2020).

Evaluating students' essays has traditionally relied on skilled human evaluators who examine content and determine quality. However, this process comes with notable limitations and challenges. Human evaluation is time-consuming and labor-intensive, particularly when the volume of essays is high. Additionally, human evaluators may introduce subjectivity and bias, leading to inconsistent results. Factors such as fatigue and boredom can further compromise accuracy, causing evaluators to overlook errors or make less precise judgments (Dikli, 2006). These challenges highlight the need for more efficient and reliable methods of essay evaluation.

Variations in human evaluations can be attributed to several factors, including the rater's experience and linguistic or cultural background. Research by Rao and Li (2017) and Barkaoui (2010a) shows that less experienced raters may adopt lenient scoring practices, while more experienced raters tend to be stricter. Similarly, cultural and linguistic differences can shape a rater's perception of what constitutes effective writing, influencing their evaluations. Studies by Barkaoui (2010b) and Attali (2016) emphasize that training and experience can significantly improve the consistency and reliability of scoring. To address these challenges, standardized rubrics, comprehensive training, and clear instructions are essential for creating a fair and objective evaluation process.

In this context, AI systems have emerged as promising tools for essay evaluation, gaining support from scholars in foreign language teaching and learning. Advocates such as Shermis and Burstein (2003) and Shermis et al. (2010) argue that AI systems are efficient, objective, and reliable. They provide consistent and immediate feedback, free from biases influenced by mood, fatigue, or external factors. AI systems can also handle large volumes of essays quickly, making them particularly useful for standardized testing and large-scale assessments. Despite these strengths, critics contend that AI lacks the human element required to assess the complexity of written work fully. For instance, AI systems may misinterpret nuanced language or fail to consider context, which could lead to errors in scoring. Nonetheless, advancements in AI technology continue to enhance their capabilities and effectiveness.

AI systems have been widely adopted in commercial testing settings, such as TOEFL, where they efficiently score essays and provide precise evaluations. Beyond standardized tests, these systems have found applications in Massive Open Online Courses (MOOCs), where they check written assignments and offer tailored feedback. This functionality not only improves students' writing skills but also reduces the workload for instructors. By providing prompt and detailed feedback, AI systems support scalable learning environments and enable greater student participation in MOOCs.

However, limitations remain, particularly regarding the applicability of AI systems to learners from diverse linguistic and cultural backgrounds. As Wali and Huijser (2018) note, many AI-based writing-enhancement systems are built using datasets that primarily represent English as used in native-speaking contexts. These systems may struggle to accommodate differences in grammar, syntax, and semantics across various English dialects or adapt to the linguistic needs of learners whose first language is not English. For instance, errors in writing conventions or tone that arise from cultural differences may not be accurately identified or addressed by AI systems, potentially leading to ineffective feedback.

Given these challenges, it is critical to supplement AI systems with human guidance, especially for non-native English speakers. Human instructors can address nuances in language use and cultural context that AI systems often overlook. Furthermore, research and development should focus on creating more inclusive and culturally sensitive AI tools to better support diverse learners. This study aims to contribute to this effort by assessing the efficiency of a recent AI system, DeepAI, which is trained by researchers to correct mistakes, score essays, and provide feedback while addressing some of the challenges associated with traditional AI systems. The research questions proposed for this study are:

1. What categories of errors are identified in the participants' writings by AI systems and human evaluators?
2. What is the correlation between the errors identified by human evaluators and the ones identified by AI systems?

Literature Review

Error Analysis (EA) process, as defined by Corder (1967), refers to a method that can be used to evaluate the linguistic performance of people who do not speak English as their mother tongue. Corder emphasized the importance of EA, while Brown (2000) described it as reflective since it could be implemented to understand the learners' knowledge of remedial methods to develop language structures. According to Corder, EA could provide guidance to instructors and syllabus designers to develop effective remedial courses for the target language. With this perspective, teachers could identify recurring errors made by learners, assess their current level of language proficiency, and use such errors as an opportunity to devise educational approaches that help improve language skills.

Other scholars (Richards & Schmidt, 2002; Richards, 1974) have classified various kinds of errors in texts, including overproduction, communication, developmental, simplification, and overgeneralization errors. Corder (1981) further explained that remedial Error Analysis was designed to help teachers evaluate and correct errors, while developmental EA emphasized the interlanguages used by language learners. Corder's Error Analysis approach (1974) was employed in multiple studies (Huang, 2001; Chastian, 1990). The primary aim of Corder's Error

Analysis approach was to inspire researchers, practitioners as well as instructors to understand how foreign language learning occurs naturally through learners' errors and to use this knowledge to develop sound pedagogical practices.

Gamper and Knapp (2002) argued that AI software has a significant role in enhancing EFL writing proficiency and evaluating and providing feedback on writing, as prior research has demonstrated (Ranalli, 2018; Li et al., 2015). Corrective feedback is a significant feature of any writing course (Chen, 2016), and explicit feedback, particularly, assists learners in comprehending effective writing, reflecting on writing skills, and improving their writing development (Ranalli, 2018).

Two scholars (Shermis & Burstein, 2003) defined AI-based evaluation systems as measurement technologies capable of emulating written production. An AI system utilizes multiple methods and other similar technologies to check the essay's structure and quality (Burstein & Chodorow, 2010). Current AI systems, as per Ke (2019), assess different qualities of a text, such as persuasiveness, clarity, coherence, cohesion, development, organization, relevance, style, mechanics, word usage, as well as grammaticality.

Numerous studies have investigated AI-based scoring during the past decades, with researchers, including Attali (2013) and Bennett & Bejar (1998), highlighting its scoring simplicity, reliability, objectivity, and consistent accuracy with human evaluation. Several studies have also revealed the positive impact such systems have had on students' writing quality as well as motivation, such as encouragement of more revisions and the production of extended texts (Li et al., 2014, 2015), as well as improvement in punctuation, spelling, grammar as well as vocabulary (Jayavalan & Razali, 2018). In addition, these systems have been compared to human rating in various studies, showing a significant level of correlation (Dikli & Bleyle, 2014).

It has been suggested that AI based scoring systems primarily examine the surface-level syntax and semantically relevant lexicon to evaluate the quality of writing, while human scores rely on other aspects of writing beyond syntax and lexicon (Huang, 2014). Huang also argued that these systems focused on general text features, such as lexical range, sentence length, and word count, while ignoring the texts’ stylistic and rhetorical features. Similarly, a scholar noted that such systems emphasized the correctness of writing and overlooked its cultural and social aspects (Vojak et al., 2011).

Another study by Dikli and Bleyle (2014) evaluated the implementation of such systems in an academic writing context, where they analyzed grammar, usage, and mechanics. The results indicated that the instructors provided rich quality and quantity feedback compared to the system, and they also identified a higher number of errors in terms of types and categories. The authors suggested that the limitations of the systems should be mentioned to both instructors and learners before they are integrated into the language-learning context. Another study by Park (2019) compared one such system's performance to human rating and found that the system failed to diagnose as many as 156 errors detected by human scorers, including errors in adverbs, auxiliary verbs, conjunctions, determinatives, mechanics, noun agreement, noun forms, prepositions, pronouns, sentence-level, tense, verb agreement, verb form, word order and wrong noun and verb choices. These findings highlight the challenges of using these systems as a sole indicator of writing quality and suggest that human raters' input remains crucial.

In a study by Dembsey (2017), the performance of an AI system was compared with that of writing experts. The study found that the system generated the highest percentage of cumulative comments compared to those of the experts. However, 47 percent of the 118 cumulative comments were concerned with the application of stylistic rules. The author came to the conclusion that although the system could be accessed anywhere and at any time, it could not provide a variety of support types, including individual feedback, praise, agency, and support on issues beyond sentences. These are aspects that can be provided during a writing consultation, where students can receive feedback on their writing holistically rather than focusing solely on grammar and stylistic rules. Therefore, while these systems can be helpful for some aspects of writing, they should be viewed as support, not replacement, for human writing evaluation.

Despite the limitations of AI systems in providing non-judgmental, contextualized, and personalized feedback, a study by Wali and Huijser (2018) expressed willingness to employ such tools. The authors recommended using a mixture of these systems and teacher feedback. These systems can also provide explicit feedback for writing, as confirmed in a study by Ranalli (2018). The study indicated that the tool creators offered either specific generic feedback based on writers’ capabilities and requirements or the writing tasks. This approach would enable learners to receive more effective feedback, enhancing their writing skills and overall writing quality. Therefore, while these systems have their limitations, they can be beneficial when combined with other feedback sources and when designed to provide specific feedback according to the learners' specific needs and tasks.

Studies of this type have shown that using AI can be useful in reducing learner errors in areas such as punctuation, spelling, syntax, and lexicon in students' writing and can promote their autonomy. Additionally, they can be helpful in improving students' sentence construction and grammar functions. However, these studies have also shown that indirect corrective feedback from teachers has a direct positive effect on both content quality and organization of students’ writing. Thus, while AI tools can be beneficial, it is still imperative to provide training and guidance to students so that they can use them effectively, particularly with low-proficiency learners.

Scholars such as Park and Yang (2020) found that AI feedback on syntax was the most frequent, and the intensive and immediate feedback helped students promote their accuracy and awareness of the use of determiners. The study suggested that AI systems could provide corrective feedback on various linguistic issues, allowing students to enhance writing skills in general, beyond just grammar and spelling. Additionally, other scholars such as Chukharev-Hudilainen and Saricaoglu (2016) prepared tools capable of discourse analysis, which was highly effective in providing feedback. They suggested that feedback on discourse could also be included in AI systems to improve their capabilities beyond just checking for grammatical and structural correctness, such as user-friendliness.

In addition, a study conducted by Wilson and Roscoe (2019) found that teachers had a positive view of AI tools' social validity, using contextual factors to evaluate their effectiveness. Similarly, another study by Lu et al. (2019) reported the positive effect of AI as a self-evaluation tool. This approach encouraged autonomous and self-regulated learning, where teachers could promote the implementation of AI systems in foreign language classes to motivate students to evaluate their writings, edit them based on AI feedback, and produce better-quality work.

Numerous studies have been conducted regarding AI systems, but researchers have also identified limitations within this field for various reasons. For instance, two researchers (Dikli & Bleyle, 2014) argued that the majority of research in the field has centered on writing by native writers in large-scale assessment centers. Furthermore, Kassim (2019) suggested that studies with small sample sizes may affect their validity and lead to untrustworthy results. In their study, Farangi and Zabbah explored the use of Artificial Neural Networks (ANNs) and Neuro-fuzzy Systems (NFS) to predict the performance of Iranian EFL (English as a Foreign Language) students in a reading comprehension course. The researchers aimed to compare the prediction accuracy of these models against the scores given by instructors. They found that both ANNs and NFS were effective in forecasting students' final scores, with NFS showing particular promise due to its ability to handle uncertainties and adapt to changing data patterns. Therefore, more research is necessary to investigate how these systems perform in diverse contexts and settings. Therefore, this study aimed to find out how an AI tool (DeepAI) performs in the EFL context in comparison to evaluations made by human rating.

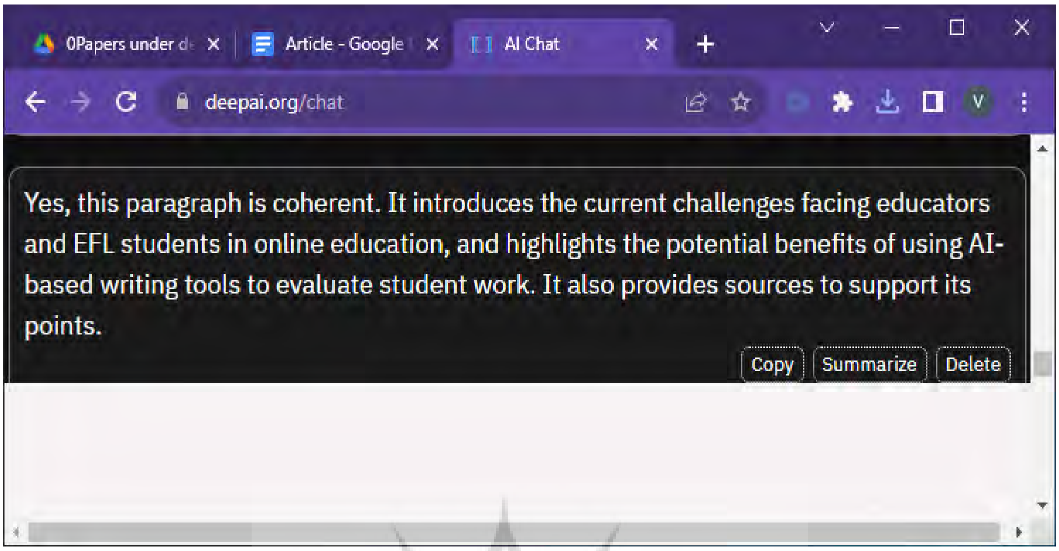
While previous studies have explored the capabilities of AI-based essay scoring systems, significant gaps remain in understanding their effectiveness in diverse linguistic and cultural contexts. Much of the existing research, such as that by Shermis and Burstein (2003) and Attali (2016), focuses on AI systems applied to standardized tests or native English-speaking populations. These studies often rely on datasets derived from Western contexts, limiting their applicability to English as a Foreign Language (EFL) learners, particularly those from non-Western backgrounds like Persian-speaking students. Such populations exhibit unique linguistic and cultural characteristics, including syntax, grammar, and stylistic conventions, which may not be adequately captured by AI systems trained on generalized datasets. Consequently, there is a need for localized evaluations of AI scoring systems to assess their adaptability and efficacy in identifying error patterns specific to these learners.

Additionally, while many studies have examined AI's ability to provide immediate, objective feedback, few have compared its error-detection capabilities to human raters across detailed categories, especially using established linguistic frameworks such as Corder's (1974) Error Analysis. Research by Dikli and Bleyle (2014) and Park (2019) has highlighted discrepancies between AI and human scoring, but these investigations often focus on broad outcomes without delving into specific error types or their implications for pedagogy. This gap is particularly relevant for EFL learners, where targeted feedback on recurring errors can significantly impact language acquisition. This study addresses these gaps by juxtaposing the performance of an AI-based system, DeepAI, with human evaluators in an EFL context, offering insights into the nuanced strengths and limitations of automated scoring systems when applied to linguistically and culturally distinct learners.

Method

Participants

This study utilized a quantitative approach, and the data was collected from a sample of 200 full-time first- and second-year EFL students through their written samples. The participants were selected from eleven classes taught by teachers holding Ph.D. degrees in Teaching English as a Foreign Language (TEFL). The writing classes were taught at a female-only state university in Tehran, where English was the main medium of instruction. DeepAI (deepai.org/chat) was used as the main AI in the study. Figure 1 displays the AI interface. As can be seen in the figure, the system has evaluated the input as coherent and has given details for the coherence.



Yes, this paragraph is coherent. It introduces the current challenges facing educators and EFL students in online education, and highlights the potential benefits of using AI-based writing tools to evaluate student work. It also provides sources to support its points.

Figure 1. *DeepAI interface*

Data Collection

For data collection, the researchers utilized convenience sampling based on recommendations by Morse (1991), including a corpus of 200 essays collected from students' exams. The students had no idea that their writings would be analyzed but agreed to allow the researchers to use exam papers for future research objectives anonymously. During these exams, the students had three optional topics related to essay writing, and they were required to write around 500 words in one hour. The exams were conducted in several computer labs with local network access, allowing students only to save their final output on the server. Each learner was provided with a laptop and a blank file in which the exam instructions and topics for the essay were already included. Students were instructed to provide one essay on one of the given topics: "Oil in the Middle East: A Fortune or a Curse?” “The importance of English writing in college survival” and “the role of self-confidence in success.”

Evaluation procedure

Two instructors graded the essays using a rubric designed and developed by the researchers. They were both PhD holders in TEFL and also had experience scoring IELTS task 2 writings. Instructors have used the rubric for several years, and it has been authenticated and examined by professionals in TEFL. The English department also endorsed it as a standard for grading essays. The rubric evaluated standards including coherence, mechanics, syntax, paragraphs, topic sentences, thesis statement, introduction, and conclusion.

Results

The study classified writing errors by using Corder's (1974) error analysis, identifying four categories of errors: textual, referential, register, and social errors. Within the context of this study, EFL learners mainly made systematic errors identified by both AI and human rating. There were fewer instances of referential errors and social errors.

The researchers and an external rater with a Ph.D. in TEFL and experience in essay scoring used rubrics to score each student's essay, although it was very hard and time-consuming to check all the essays. Cohen's kappa (1960) was implemented to calculate the interrater reliability between the scores. The calculated kappa values ranged from 0.693 to 0.944. Therefore, the interrater reliability of scoring between the external rater and the researchers was deemed acceptable.

Table 1 presents the total number of errors, mean of errors, and Standard Deviation (SD) for each type of error in essays for both human rating and AI across all 200 essays. Furthermore, the table includes a Pearson correlation coefficient analysis that shows the relationship between the errors as identified by AI and human rating.

Table 1.

Total number, mean number, and SD for each error type

Error Type	**Human**	**Mean**	**SD**	**AI**	**Mean**	**SD**	**r**	**p**
Verb form	118	0.6	1.16	146	0.75	1.37	0.598	0
Subject verb agreement	856	4.35	4.77	975	4.95	3.75	0.363	0
tense	185	0.94	1.24	367	1.87	2.26	0.437	0
determiner	127	0.65	1.26	115	0.59	1.52	0.545	0
preposition	424	2.15	2.09	412	7.17	3.77	0.371	0
Word order	43	0.22	0.51	2	0.02	0.08	0.114	0.117
collocation	380	1.94	2.61	752	3.82	3.02	0.243	0.001
idiom	15	0.07	0.31	1	0.03	0.29	-0.016	0.821
Word choice	273	1.39	1.78	405	2.03	2.49	0.709	0
spelling	16	0.09	0.35	76	0.39	0.81	0.026	0.731
Punctuation	282	1.44	2.1	403	2.05	2.38	0.744	0
capitalization	49	0.25	0.71	1	0.01	0.01	-	–
Contraction	82	0.42	0.79	1	0.01	0.01	-	-
Passive	15	0.08	0.35	192	1.01	1.27	0.172	0.017
style	81	0.41	1.25	1	0.01	0.01	-	–
Sentence structure	20	0.11	0.34	144	0.74	1.02	-0.09	0.216
dangling	72	0.37	0.75	134	0.69	1.09	0.333	0
Comma splice	125	0.63	1.22	2010	10.21	5.48	-0.235	0.001
coherence	1	0.02	0.07	352	1.79	1.72	0.053	0.478
cohesion	287	1.46	2.05	103	0.53	0.99	-0.054	0.46

The analysis revealed differences in the number and types of errors identified by AI and human raters across various categories. AI generally identified a higher frequency of errors in categories such as subject-verb agreement, verb forms, and tense compared to human raters. However, the data did not provide conclusive evidence to suggest AI's superiority across all error categories.

A Pearson correlation analysis showed a strong positive correlation between the total error frequencies identified by AI and human raters, indicating general agreement in their error-detection trends. When analyzing specific error types, the recalculated correlations revealed moderate to strong positive correlations in categories like verb forms and subject-verb agreement. This suggests some alignment in the accuracy of the two approaches for these error types.

For other categories, such as noun numbers, passive voice misuse, and prepositions, the correlation was weaker, suggesting some divergence between the two methods in error identification. Moreover, in certain categories like improper formatting, contractions, and informal language, there was no significant correlation, indicating inconsistent error detection.

Interestingly, while the description highlighted "negative correlations" for lexical selection, contraction errors, and informal language, the recalculated data did not support these claims. Instead, these categories showed either no significant correlation or insufficient data for statistical analysis.

In summary, the recalculated data indicate a generally positive relationship between AI and human error detection, with variations in agreement depending on the error category. The results suggest that while AI and human raters often identify similar patterns of errors, there remain significant differences in certain areas that warrant further investigation.

Paired sample T-test

Table 2 presents the results of paired-sample t-tests to compare the scores by AI and human rating. The results revealed significant differences between the total grades proposed by the human rating and AI, with a p-value less than .05. The reported mean difference suggested that the overall score as given by human rating was significantly higher compared to that suggested by AI. Additionally, a significant difference was found between the sum of the errors diagnosed by human rating and AI, with a p-value of less than .05. This difference indicated that the sum of errors diagnosed by human raters was less than that derived using AI, implying that AI detected higher numbers of errors as compared with human rating. This result can be attributed to the p-values being less than .05, and the negative mean value. Thus, AI tended to diagnose significantly more errors on these error categories than human rating.

Table 2.
Paired-sample t-test Result

Error type	Mean difference	Std deviation	t-test	p-value
Verb form	-0.16245	1.14466	-1.993	0.044
Subject verb agreement	-0.60407	4.87347	-1.741	0.084
Tense	-0.92387	2.03521	-6.372	0.000
Determiner	0.06092	1.33492	0.641	0.523
Preposition	-5.01524	3.56497	-9.747	0.000
Word order	0.20919	0.79382	5.878	0.000
Collocation	-1.88326	3.46876	-7.621	0.000
Idiom	0.05077	0.42555	1.675	0.096
Word choice	-0.63453	1.74919	-5.092	0.000
Spelling	-0.30458	0.85634	-4.993	0.000
Punctuation	-0.61422	1.61733	-5.331	0.000
Capitalization	0.24874	0.71004	4.918	0.000
Contraction	0.41118	0.78148	7.386	0.000
Passive	-0.89849	1.24535	-0.127	0.000
Style	0.40640	1.24030	4.596	0.000
Sentence structure	-0.62945	1.09254	-8.087	0.000
Dangling	-0.31473	1.08912	-4.057	0.000
Comma splice	-9.57869	5.87630	-2.880	0.000
Coherence	-1.77666	1.70859	-4.596	0.000
Cohesion	0.93910	2.30495	5.720	0.000

In contrast, significant variances between the two human scoring and AI were found in the sum of errors pertaining to a comma splice, run-on sentence, possessive, and coordinating conjunction, as indicated by p-values less than 0.05 and positive mean values. The findings suggested that human rating detected significantly more errors in these categories than AI.

Discussion

This study explored the similarities and differences between an AI-based scoring system (DeepAI) and human raters in identifying errors in the essays of Persian-speaking EFL learners. The findings provide valuable insights into the capabilities and limitations of AI systems, highlighting areas where they align with or diverge from human evaluation.

1. Comparison of AI and Human in Error Identification

The results showed notable differences in the types and frequencies of errors identified by AI and human raters. AI demonstrated superior performance in detecting surface-level errors, such as punctuation, prepositions, verb forms, and lexical choices. This aligns with findings from Park (2019) and Dikli and Bleyle (2014), who reported that AI systems excel in areas where grammatical and syntactic rules are explicit and quantifiable. The ability of AI to process large datasets and apply consistent rules likely accounts for its higher accuracy in these categories.

Conversely, human raters outperformed AI in identifying more nuanced and context-dependent errors, including issues with sentence structure, coherence, and stylistic appropriateness. This is consistent with studies by Vojak et al. (2011) and Dembsey (2017), which highlighted the limitations of AI systems in interpreting rhetorical and stylistic elements that require a deeper understanding of context and intent. For instance, human raters were more adept at recognizing errors in run-on sentences and comma splices—areas where AI struggled due to its reliance on surface-level analysis.

2. Correlations Between AI and Human Scores

The moderate correlation found between AI and human scores supports the notion that AI systems can partially replicate human judgment, particularly in quantitative aspects of error detection. Previous research by Shermis and Burstein (2003) also found a significant overlap between AI and human scoring in detecting mechanical and grammatical errors. However, discrepancies in error categories such as coherence and cohesion suggest that AI systems, despite their advancements, cannot fully replicate the qualitative assessment provided by human raters.

The differences in error detection can be attributed to the inherent strengths and limitations of AI systems and human raters. AI's reliance on rule-based algorithms and training data makes it highly effective for detecting frequent and standardized error patterns. In contrast, human evaluators bring a nuanced understanding of language that enables them to identify complex errors influenced by cultural, rhetorical, or contextual factors. For example, errors related to cohesion or idiomatic expressions are often subjective and context-sensitive, which AI systems, as noted by Huang (2014), tend to overlook.

Furthermore, the finding that AI identified a higher total number of errors compared to human raters can be justified by its tendency to flag even minor infractions consistently. While this can be advantageous for providing comprehensive feedback, it may also lead to overcorrection, as highlighted by Dembsey (2017). On the other hand, human raters are more selective, focusing on significant errors that impact meaning or clarity, which may explain their lower error counts.

These findings align with earlier studies emphasizing the complementary nature of AI and human evaluation. Ranalli (2018) and Li et al. (2014) advocate for integrating AI tools into writing assessments to enhance efficiency and provide detailed feedback while also emphasizing the irreplaceable role of human raters in offering contextualized and personalized insights. The current study corroborates this dual approach, suggesting that AI is best utilized as a supplementary tool rather than a standalone evaluator.

Conclusion

This study demonstrated that AI, although not specifically designed as a language teaching tool, can be beneficial to the language teaching community in general and writing instructors in particular. Therefore, it cannot outperform human raters simply by taking into account the number of errors identified. In spite of their high accuracy, at times, AI identifies correct structures as errors, especially when the structures are complex or may have ambiguous structures. Therefore, too much reliance on the tool should not be encouraged.

The findings of this study have implications for EFL writing education. Incorporating AI systems into EFL writing instruction can enhance the objectivity and consistency of writing assessment. AI systems can provide prompt feedback to EFL learners, enabling them to diagnose and correct errors in their writing autonomously. AI systems can be used in conjunction with classroom instruction to promote language awareness and metalinguistic skills among EFL learners.

Incorporating AI systems into EFL writing education enhances consistency by eliminating the variability often associated with human raters. Human evaluators, influenced by factors such as fatigue, mood, cultural background, or personal biases, may score the same piece of writing differently. In contrast, AI systems apply standardized algorithms and predetermined rules uniformly, ensuring that every essay is assessed against the same criteria without subjective influence. This uniformity allows for more reliable comparisons across students' performances, fostering fairness in assessment. Additionally, AI's ability to consistently detect specific error types—such as grammar, punctuation, and syntax errors—means learners receive uniform feedback, which helps them focus on recurring issues. This consistency not only benefits large-scale assessments but also supports individualized learning by providing EFL learners with predictable, objective evaluations that promote gradual improvement.

EFL instructors can encourage the implementation of AI as an assessment and assistance tool to check and aid writing after brainstorming and during the final phase of writing. AI can also support writing teachers to assess students' work more efficiently, allowing them to identify possible problematic areas before submitting the final version to the teacher, thus promoting self-assessment and motivational strategies.

It is important to be aware of AI's weaknesses and strengths and to use it judiciously. While AI can detect many errors and offer valuable suggestions, it cannot detect all errors. Therefore, EFL instructors should guide learners on how to use AI effectively, deal with errors beyond the basic ones identified by AI, and use AI as a complementary tool to receive differentiated feedback. In this sense, AI can be a valuable tool for EFL students and instructors alike, providing comprehensive and personalized feedback tailored to their specific needs and capabilities.

Despite the positive findings, this study had some limitations that can be removed in future research. The sample was limited to EFL learners within a university context, which may limit generalizations about the usefulness of AI in other EFL contexts. More diverse samples of EFL learners are needed to confirm the results and broader applicability of AI systems. The sample size was relatively small, possibly affecting the statistical generalization of the study and limiting its generalizability to larger populations. Future studies should consider expanding the sample size to help confirm the results. This study was limited to a single writing course, which may not represent the whole population of students in writing courses. Future studies should investigate the use of AI across different courses and proficiency levels to help generalize the results.

It asl to be noted that AI is not necessarily accurate in determining these types of errors. Normally, AI systems are not perfectly accurate, so just focusing on the number of errors would be misleading. What is important is whether they identify the same types of errors. Since the error types have not been compared, this can also be considered as a serious limitation of this study.

Acknowledgments

I would like to thank the editorial team of TESL Quarterly for granting me the opportunity to submit and publish the current synthesis. I would also like to express my appreciation to the anonymous reviewers for their careful, detailed reading of my manuscript and their many insightful comments and suggestions.

Declaration of conflicting interests

The author declares no potential conflicts of interest concerning the research, authorship, and/or publication of this article.

Funding

The author received no financial support for this article's research, authorship, and/or publication.

References

Al-Ahdal, A. (2020). Using computer software as a tool of error analysis: Giving EFL teachers and learners a much-needed impetus. *International Journal of Innovation, Creativity, and Change, 12*(2), 418–437. https://doi.org/10.3390/languages4010019

Alrashidi, O., & Phan, H. (2015). Education context and English teaching and learning in the Kingdom of Saudi Arabia: An overview. *English Language Teaching, 8*(5), 33–44. https://doi.org/10.5539/elt.v8n5p33

Alshakhi, A. (2019). Revisiting the writing assessment process at a Saudi English language institute: Problems and solutions. *English Language Teaching, 12*(1), 176–185. https://doi.org/10.5539/elt.v12n1p176

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

Attali, Y. (2016). A comparison of newly-trained and experienced raters on a standardized writing assessment. *Language Testing, 33*(1), 99–115. https://doi.org/10.1177/0265532215582283

Attali, Y., Lewis, W., & Steier, M. (2013). Scoring with the computer: Alternative procedures for improving the reliability of holistic essay scoring. *Language Testing, 30*(1), 125–141. https://doi.org/10.1177/0265532212452396

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. *The Journal of Technology, Learning and Assessment, 4*(3).

Barkaoui, K. (2010a). Do ESL essay raters' evaluation criteria change with experience? A mixed-methods, cross-sectional study. *TESOL Quarterly, 44*(1), 31–57. https://doi.org/10.5054/tq.2010.214047

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

Bennett, R. E., & Bejar, I. I. (1998). Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practice, 17*, 9–17. https://doi.org/10.1002/j.2333-8504.1997.tb01734.x

Burstein, J., & Chodorow, M. (2010). Progress and new directions in technology for automated essay evaluation. In R. B. Kaplan (Ed.), *The Oxford handbook of applied linguistics* (pp. 529–539). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195384253.013.0036

Burstein, J., Marcu, D., & Knight, K. (1998). A machine learning approach to recognizing features of coherence in student essays. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*.

Chen, C., Cheng, Y., & Huang, H. (2020). The impact of automated feedback on EFL learners' writing performance: A meta-analysis. *Educational Technology Research and Development, 68*(2), 123-145.

Chukharev-Hudilainen, E., & Saricaoglu, A. (2016). Causal discourse analyzer: Improving automated feedback on academic ESL writing. *Computer Assisted Language Learning, 29*(3), 494–516. https://doi.org/10.1080/09588221.2014.991795

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Corder, S. P. (1967). The significance of learners' errors. *International Review of Applied Linguistics, 5*, 161–170.

Corder, S. P. (1981). *Error analysis and interlanguage*. Oxford University Press.

Cotos, E. (2015). Automated writing analysis for writing pedagogy. *Writing & Pedagogy, 7*(2–3), 197–231. https://doi.org/10.1558/wap.v7i2-3.26381

Dembsey, J. M. (2017). Closing the Grammarly® gaps: A study of claims and feedback from an online grammar program. *The Writing Center Journal*, *36*(1), 63–96. https://www.jstor.org/stable/44252638

Dikli, S., & Bleyle, S. (2014). Automated essay scoring feedback for second language writers: How does it compare to instructor feedback? *Assessing Writing, 22*, 1–17. https://doi.org/10.1016/j.asw.2014.03.006

Douglas, D. (2010). *Understanding language testing*. Hodder Education.

Farangi, M. R., & Zabbah, M. (2023). Intelligent scoring in an English reading comprehension course using artificial neural networks and neuro-fuzzy systems. *Teaching English as a Second Language Quarterly*, 42(4), 1-21. Retrieved from https://tesl.shirazu.ac.ir

Gamper, J., & Knapp, J. (2002). A review of intelligent CALL systems. *Computer Assisted Language Learning*, *15*(4), 329–342. https://doi.org/10.1076/call.15.4.329.8270

Ghufron, M. A., & Rosyida, F. (2018). The role of Grammarly in assessing English as a foreign language (EFL) writing. *Lingua Cultura*, *12*(4), 395–403. https://doi.org/10.21512/lc.v12i4.4582

Goh, T. T., Sun, H., & Yang, B. (2020). Microfeatures influencing writing quality: The case of Chinese students' SAT essays. *Computer Assisted Language Learning*, *33*(4), 455–481. https://doi.org/10.1080/09588221.2019.1572017

Gonzalez, M., Liu, Y., & Zhang, J. (2021). Addressing bias in automated essay scoring: A case study on EFL learners' essays. *Language Testing*, *38*(4), 495-515.

Higgins, J. J. (1983). Computer-assisted language learning. *Language Teaching*, *16*(2), 102–114.

Huang, S. J. (2001). Error analysis and teaching composition [Unpublished master's thesis]. National Tsing Hua University.

Huang, S. J. (2014). Automated versus human scoring: A case study in an EFL context. *Electronic Journal of Foreign Language Teaching*, *11*, 149–164. https://doi.org/10.1080/15434303.2016.1230121

Jayavalan, K., & Razali, A. B. (2018). Effectiveness of online grammar checkers to improve secondary students' English narrative essay writing. *International Research Journal of Education and Sciences*, *2*(1), 1–6. http://psasir.upm.edu.my/id/eprint/14442

Ke, Z. (2019). Automated essay scoring: A survey of the state of the art [Paper presentation]. *International Joint Conference on Artificial Intelligence 28th Annual Meeting*. https://doi.org/10.24963/ijcai.2019/879

Kenning, M. J., & Kenning, M. M. (1983). *Introduction to computer-assisted language teaching*. Oxford University Press.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated essay assessment. *Assessment & Evaluation in Higher Education*, *28*(5), 491-505.

Li, J., Link, S., & Hegelheimer, V. (2015). Rethinking the role of automated writing evaluation (AWE) feedback in ESL writing instruction. *Journal of Second Language Writing, 27*, 1–18. https://doi.org/10.1016/j.jslw.2014.10.004

Li, Z., Link, S., Ma, H., Yang, H., & Hegelheimer, V. (2014). The role of automated writing evaluation holistic scores in the ESL classroom. *System*, *44*, 66–78. https://doi.org/10.1016/j.system.2014.02.007

Lu, M., Deng, Q., & Yang, M. (2019). EFL writing assessment: Peer assessment vs. automated essay scoring. In E. Popescu, H. Tianyong, T.-C. Hsu, H. Xie, & M. Temperini (Eds.), *International*

Symposium on Emerging Technologies for Education (pp. 21–29). Springer. https://doi.org/10.1007/978-3-030-38778-5_3

Morse, J. M. (1991). Strategies for sampling. In J. M. Morse (Ed.), *Qualitative nursing research: A contemporary dialogue* (pp. 127–145).

Nova, M. (2018). Utilizing Grammarly in evaluating academic writing: A narrative research on EFL students' experience. *Premise: Journal of English Education*, *7*(1), 80–97. https://doi.org/10.24127/pj.v7i1.1300

O'Neill, R., & Russell, A. M. T. (2019). Stop! Grammar time: University students' perceptions of the automated feedback program Grammarly. *Australasian Journal of Educational Technology*, *35*(1), 42–56. https://doi.org/10.14742/ajet.3795

Park, J. (2019). An AI-based English grammar checker vs. human raters in evaluating EFL learners' writing. *Multimedia-Assisted Language Learning*, *22*(1), 112–131. https://doi.org/10.15702/mall.2019.22.1.112

Perdana, I., & Farida, M. (2019). Online grammar checkers and their use for EFL writing. *Journal of English Teaching, Applied Linguistics, and Literatures*, *2*(2), 67–76. https://doi.org/10.20527/jetall.v2i2.7332

Prinsloo, D., & Bothma, T. (2020). A copulative decision tree as a writing tool for Sepedi. *South African Journal of African Languages*, *40*(1), 85–97. https://doi.org/10.1080/02572117.2020.1733834

Polit, D. F., & Hungler, B. P. (1993). *Study guide for essentials of nursing research: Methods, appraisal, and utilization*. Lippincott, Williams, & Wilkins.

Ranalli, J. (2018). Automated written corrective feedback: How well can students make use of it? *Computer Assisted Language Learning*, *31*(7), 653–674. https://doi.org/10.1080/09588221.2018.1428994

Rao, Z., & Li, X. (2017). Native and non-native teachers' perceptions of error gravity: The effects of cultural and educational factors. *The Asia-Pacific Education Researcher*, *26*(1–2), 51–59. https://doi.org/10.1007/s40299-017-0326-5

Reilly, E. D., Stafford, R. E., Williams, K. M., & Corliss, S. B. (2014). Evaluating the validity and applicability of automated essay scoring in two massive open online courses. *International Review of Research in Open and Distributed Learning*, *15*(5), 83–98. https://doi.org/10.19173/irrodl.v15i5.1857

Seker, M. (2018). Intervention in teachers' differential scoring judgments in assessing L2 writing through communities of assessment practice. *Studies in Educational Evaluation*, *59*, 209–217. https://doi.org/10.1016/j.stueduc.2018.08.003

Sharma, C., Bishnoi, A., Sachan, A. K., & Verma, A. (2019). Automated essay evaluation using natural language processing. *International Research Journal of Engineering and Technology*, *5*(6), 2055–2058. https://www.irjet.net/archives/V6/i5/IRJET-V6I5398.pdf

Shermis, M., & Burstein, J. (2003). *Automated essay scoring: A cross-disciplinary perspective*. Routledge.

Shermis, M. D., & Hamner, B. (2012). Contrasting state-of-the-art essay scorers: A preliminary report on the first automated essay scoring challenge. *Proceedings of the 2012 Conference on Computer-Based Test Development*.

Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing, 20*, 53–76. https://doi.org/10.1016/j.asw.2013.04.001

Shermis, M. D., Burstein, J., Higgins, D., & Zechner, K. (2010). Automated essay scoring: Writing assessment and instruction. In E. Baker, B. McGaw, & N. S. Petersen (Eds.), *International encyclopedia of education* (3rd ed., pp. 75–80). Elsevier.

Sparks, J. R., Song, Y., Brantley, W., & Liu, O. L. (2014). Assessing written communication in higher education: Review and recommendations for next-generation assessment (Issue No. 2). *ETS Research Report Series*. https://doi.org/10.1002/ets2.12035

Vojak, C., Kline, S., Cope, B., McCarthey, S., & Kalantzis, M. (2011). New spaces and old places: An analysis of writing assessment software. *Computers and Composition*, *28*(2), 97–111. https://doi.org/10.1016/j.compcom.2011.04.004

Wali, F. A., & Huijser, H. (2018). Write to improve: Exploring the impact of an automated feedback tool on Bahraini learners of English. *Learning & Teaching in Higher Education: Gulf Perspectives*, *15*(1). https://doi.org/10.18538/lthe.v15.n1.293

Wang, J., & Brown, M. S. (2007). Automated essay scoring versus human scoring: A comparative study. *Journal of Technology, Learning, and Assessment*, *6*(2). https://files.eric.ed.gov/fulltext/EJ838612.pdf

Wilson, J., & Roscoe, R. (2019). Automated writing evaluation and feedback: Multiple metrics of efficacy. *Journal of Educational Computing Research*, *58*(1), 87–125. https://doi.org/10.1177/0735633119830764

Wiseman, C. S. (2012). Rater effects: Ego engagement in rater decision-making. *Assessing Writing*, *17*(3), 150–173. https://doi.org/10.1016/j.asw.2011.12.001

Wind, S. A., & Engelhard, G. Jr. (2013). How invariant and accurate are domain ratings in writing assessment? *Assessing Writing*, *18*(4), 278–299. https://doi.org/10.1016/j.asw.2013.09.002

Wu, H., & Garza, E. V. (2014). Types and attributes of English writing errors in the EFL context: A study of error analysis. *Journal of Language Teaching & Research*, *5*(6), 125–141. https://doi.org/10.4304/jltr.5.6.1256-1262

Zhang, Y., Wang, L., & Li, X. (2019). The effectiveness of automated essay scoring: A systematic review and meta-analysis. *Computers & Education*, *139*(1), 56-68.

Appendix
Rubrics to score students' essays

Here's a comprehensive essay scoring rubric designed to assess various aspects of students' written essays, aligned with academic standards. Each criterion is scored on a 5-point scale, where 5 represents excellent performance and 1 represents poor performance.

Essay Scoring Rubric

1. Content and Development (20%)

Focus: Does the essay effectively address the prompt with clear and well-developed ideas?

- 5 (Excellent): Ideas are insightful, thoroughly developed, and directly address the prompt. Supporting details are specific, relevant, and effectively enhance the argument or narrative.
- 4 (Good): Ideas are clear and well-developed but may lack depth or complexity. Supporting details are generally relevant but not always specific.
- 3 (Satisfactory): Ideas address the prompt but lack depth or thorough development. Supporting details are present but may be superficial or occasionally off-topic.
- 2 (Needs Improvement): Ideas are underdeveloped or only partially address the prompt. Supporting details are sparse, vague, or irrelevant.
- 1 (Poor): Ideas are minimal, off-topic, or incoherent, with no meaningful development or support.

---

2. Organization and Coherence (20%)

Focus: Is the essay logically structured with a clear introduction, body, and conclusion?

- 5 (Excellent): Essay is exceptionally organized with a clear introduction, well-connected paragraphs, and a strong conclusion. Transitions between ideas are seamless.
- 4 (Good): Essay is well-organized, with a clear structure and logical flow. Some transitions may lack fluidity.
- 3 (Satisfactory): Essay has a basic structure but lacks strong transitions or logical connections between ideas.
- 2 (Needs Improvement): Essay is poorly organized, with ideas presented in an illogical or disjointed manner. Transitions are weak or missing.
- 1 (Poor): Essay lacks a coherent structure, making it difficult to follow the argument or narrative.

---

3. Language Use and Style (20%)

Focus: Is the language appropriate, varied, and effective?

- 5 (Excellent): Language is sophisticated, precise, and varied. Word choice enhances meaning, and sentences are well-constructed and engaging.
- 4 (Good): Language is clear and appropriate, with some variety in sentence structure. Word choice is accurate but may lack sophistication.
- 3 (Satisfactory): Language is functional but lacks variety or precision. Sentences are simple, and word choice is basic.
- 2 (Needs Improvement): Language is awkward or repetitive. Word choice is imprecise, and sentences are poorly constructed.
- 1 (Poor): Language is unclear, inappropriate, or riddled with errors that impede meaning.

---

4. Grammar and Mechanics (20%)

Focus: Are grammar, spelling, and punctuation used correctly?

- 5 (Excellent): Essay contains no significant errors in grammar, spelling, or punctuation. Usage enhances readability.
- 4 (Good): Minor errors in grammar, spelling, or punctuation are present but do not detract from readability.
- 3 (Satisfactory): Several errors in grammar, spelling, or punctuation are noticeable but do not significantly impede understanding.
- 2 (Needs Improvement): Frequent errors in grammar, spelling, or punctuation interfere with readability.
- 1 (Poor): Pervasive errors in grammar, spelling, or punctuation make the essay difficult to understand.

---

5. Thesis Statement and Topic Sentences (20%)

Focus: Are the thesis statement and topic sentences clear and effective?

- 5 (Excellent): Thesis statement is clear, specific, and effectively guides the essay. Topic sentences are focused and directly support the thesis.
- 4 (Good): Thesis statement is clear but may lack specificity. Topic sentences generally support the thesis but may lack clarity.
- 3 (Satisfactory): Thesis statement is present but vague or overly broad. Topic sentences are inconsistent in their relevance or clarity.
- 2 (Needs Improvement): Thesis statement is unclear, overly general, or missing. Topic sentences are weak or do not align with the thesis.
- 1 (Poor): Thesis statement and topic sentences are absent or ineffective.

Scoring Instructions

- Assign a score (1–5) for each category.

- Multiply each category score by 20% (or adjust weighting if certain aspects are prioritized).
- Add the totals for a final score out of 100%.

Example Calculation:

- Content and Development: 4 × 20 = 80
- Organization and Coherence: 3 × 20 = 60
- Language Use and Style: 4 × 20 = 80
- Grammar and Mechanics: 3 × 20 = 60
- Thesis and Topic Sentences: 4 × 20 = 80

Total Score: (80 + 60 + 80 + 60 + 80) ÷ 5 = 72%