Journal of English Language Teaching and Learning

University of Tabriz

Volume 17, Issue 35, 2025

Combined Effects of Task Sequencing and Corrective Feedback on EFL Learners' Writing: a comparison between human raters and ChatGPT

Sara Ziaei

PhD, English Department, University of Isfahan, Isfahan, Iran. E-mail: saraziaei@fgn.ui.ac.ir

Mansoor Tavakoli (Corresponding Author)

Professor, English Department, University of Isfahan, Isfahan, Iran. E-mail: tavakoli@fgn.ui.ac.ir

ARTICLE INFO:

Received date:
2025.01.02
Accepted date:
2025.03.19

Print ISSN: 2251-7995
Online ISSN: 2676-6876

Keywords:

ChatGPT, Human Raters, Correlation, Corrective Feedback, Automated Essay Scoring

Abstract

The study, which has been derived from a larger project, examined how effective ChatGPT, compared to human raters, is for scoring writing tasks when tasks were arranged from simple to complex or vice versa. In so doing, a correlational design was employed. The participants were 113 EFL learners. Two sets of writing tasks were customized based on the SSARC (simplify, stabilize, automatize, reconstruct, complexify) model. The participants were divided into two groups. They took a pre-test and did tasks in two different orders. The tasks were rectified by the researcher and returned to them later. The participants enhanced their text based on comments on tasks. After that, they took a posttest. Human raters and ChatGPT scored the pretests and posttests. A Pearson Correlation test was run to obtain the correlation between a human rater and ChatGPT. The results indicated a strong positive correlation between scores assessed by human raters and those by ChatGPT when tasks were arranged from simple to complex ($r = 968$, $p > 05$) or complex to simple ($r = 860$, $p > 05$). These findings suggest that ChatGPT can be an effective tool for writing assessments. Suggestions for further research are discussed.

Citation: Ziaei, S & Tavakoli, M. (2025). Combined Effects of Task Sequencing and Corrective Feedback on EFL Learners' Writing: a comparison between human raters and ChatGPT. *Journal of English Language Teaching and Learning*, 17 (35), 439-452. DOI: 10.22034/elt.2025.65281.2735

* Excerpted from a PhD dissertation entitled "*Combined Effects of Task Sequencing and Corrective Feedback on EFL Learners' Writing: a comparison between human raters and ChatGPT*", Supervisor: Dr. Mansoor Tavakoli, University of Isfahan, Isfahan, Iran.

Introduction

Among the four language skills in learning English, writing in English is complicated and problematic because it is not in their native language, and they have a limited chance to write in English. Writing also produces, orders, and interprets ideas into a legible text. Consequently, it seems usual that many learners, especially foreign language learners, have trouble writing (Rattanadilok et al. 2015).

Examinees face challenges in the writing section of the International English Language Testing System (IELTS). Many individuals apply for higher education in countries where English is the primary language of instruction. The IELTS exam will assess the test takers' academic language proficiency for commencing their studies. Test takers consider writing a challenging part of the IELTS exam. Consequently, IELTS test takers must enhance their writing skills to understand how to develop ideas and paragraphs in their essays. Challenges in writing may arise from aspects such as language usage or word choice.

One way to reduce writing errors is to provide participants with corrective feedback. Despite the extensive research on written corrective feedback, it continues to be a highly contested topic in second-language writing research (Tang & Liu, 2018). The persistent debate surrounding the efficacy of corrective feedback has been linked to methodological concerns and inconsistencies in corrective feedback research (Liu & Brown, 2015).

A strategic approach to addressing writing difficulties involves sequencing writing tasks methodically—a fundamental yet challenging aspect of pedagogical design (Baralt et al., 2014). Early task-based language teaching (TBLT) frameworks proposed organizing tasks by different factors, though these models were later critiqued for relying too heavily on teacher intuition to assess these variables (Robinson, 2007). Recent research suggests that adjusting task sequences can enhance writing outcomes (Allaw & McDonough, 2019), pointing to the need for more evidence-based sequencing principles.

Among the variables that proved to be influenced by task sequencing, accuracy could be the most perceivable by the learners, because they consider a score as a result of their work minus their errors. One way to make language learners aware of their errors is to address them. Such addressing is called corrective feedback and could be provided via different means based on the task purpose. In fluency-based tasks, for example, the learner had better not be interrupted during production time and receive corrective feedback after finishing the activity. However, in accuracy-based tasks, especially when the erroneous production is the lesson's focus, the learner should be corrected on the spot. Whether to provide the correct form is another concern in corrective feedback provision. When the right form is given to the learners, the corrective feedback is called direct, and when the erroneous part is spotted without revealing the correct form, it is called indirect corrective feedback (Nguyen & Nguyen, 2022).

Automated Essay Scoring (AES) refers to an assessment system that utilizes computer technology to evaluate and grade student writing automatically by examining pertinent characteristics. Automated Essay Scoring (AES) systems have become indispensable in assessing and grading written assignments in second language (L2) contexts by integrating

techniques from linguistics, psychometrics, cognitive science, and computer science (Cotos, 2014; Lagakis & Demetriadis, 2021; Lu & Hu, 2022). The introduction of ChatGPT, a user-friendly and accessible online AI tool, has expanded possibilities for local testing environments to develop their own AES systems. Before ChatGPT's public launch in 2022, AES systems largely relied on machine learning and deep learning models, which required advanced programming expertise. ChatGPT's arrival has opened doors for addressing the unique AES needs of localized contexts.

With AI becoming more accessible, leveraging it for AES presents opportunities to ease test administrators' workload, reduce labor costs, and accelerate grading processes. However, despite advancements in AES and AI, research on the effectiveness of GPT models in AES remains limited. While mainly viewed as a language generation instrument, studies have demonstrated ChatGPT's potential for AES (Mizumoto & Eguchi, 2023) and its capacity to provide feedback on student essays. As a result, to address this research gap, the study compared the human rater with ChatGPT for writing scoring with particular task sequencing and corrective feedback.

Literature Review

Tasks are the basic unit of the syllabus in a task-based syllabus (Baralt et al., 2014) and are regarded differently by various learners. Because the primary task sequencing took place based on subjective indices (Robinson, 2007), scholars failed to conduct studies. This lack was partially removed by the proposal of the cognition hypothesis (Robinson, 2001), which emphasized simple to complex task sequencing.

In another hypothesis, the Triadic Componential Framework, details concerning task complexity, which was the task sequencing criterion, were disclosed (Robinson, 2001). Consequently, resource directing (R-dir) and resource dispersing (R-dis) factors were regarded as items determining task complexity. The former focuses students' attention on the linguistic forms. The latter, however, do not focus the learners' attention on linguistic aspects (Tajeddin& Bahador, 2012). The so-called elements paved the ground for the suggestion of the SSARC (stabilize, simplify, Automatize, reconstruct, complexify) model.

The SSARC model is a three-step guide that centers on modifying task complexity in terms of R-dir and R-dis factors. The pivoting concept in this model is the interlanguage, in that the model tries to adapt a learner's interlanguage to the recently learned items. This goal is supposed to be reached by gradually boosting the complexity of the task. In the first level, as the name suggests, the item is simplified so that it can be later stabilized in the learners' interlanguage. Thus, the task is simple in terms of both R-dir and R-dis features. In the second level, to automatize the learner's access to the interlanguage, the task becomes complex only in R-dis features. In the final level, the task is made complex on both R-dir and R-dis features to complexify the interlanguage and reconstruct it based on the recent item.

The SSARC model presents a structured three-phase approach to adjusting task complexity through the strategic manipulation of resource-directing (R-dir) and resource-dispersing (R-dis) cognitive factors, with the fundamental aim of fostering interlanguage development. At its core, the model operates on the principle of gradually adapting task demands to align with and

challenge a learner's evolving linguistic system. The initial phase simplifies tasks along both R-dir and R-dis dimensions to promote successful item integration and stabilization within the interlanguage. Subsequently, the model introduces increased R-dis complexity while maintaining simplified R-dir features, thereby encouraging automatic retrieval and fluent use of the acquired knowledge. Finally, the model simultaneously elevates complexity across both R-dir and R-dis parameters to push learners toward comprehensive interlanguage restructuring and expansion, effectively incorporating new linguistic elements into their developing system. This graduated approach systematically bridges the gap between current and target linguistic competence through carefully sequenced cognitive challenges.

Corrective feedback has long been used to enhance writing in language classes (Allaw & Mc Donough, 2019). After the controversy over its usefulness in language learning, corrective feedback received little criticism. Several studies have examined it in a task-based context (Zhai& Gao, 2018). Corrective feedback could highlight the wrong part without giving the right form, and vice versa. In case the correct form is not given to the learner, the corrective feedback is called indirect (Nguyen & Nguyen, 2022). This type of corrective feedback could solely spot the error or supply hints to the correct form. The hints could be explanations about the nature of the error or some conventional code functioning as explanations.

In the year following ChatGPT's release, numerous articles have explored its potential as an assessment tool (Nam& Bai, 2023; Poole& Coss, 2024). Although primarily recognized as a language creation platform, research has highlighted its promise in automated essay scoring (AES) (Mizumoto & Eguchi, 2023) and providing feedback on student writing. Despite these developments, the use of AI for evaluating and scoring student exam work remains contentious (Nam & Bai, 2023). Nevertheless, ChatGPT’s prompt-based interface has made AI accessible to a wide audience, including language testing institutions. Developing reliable methods to leverage AI for AES could benefit test administrators by reducing workload, lowering labor costs, and improving turnaround time.

Poole and Coss (2024) investigated the effectiveness of ChatGPT in assessing L2 writing, focusing on its feasibility as an evaluation tool. The research analyzed the accuracy and reliability of AI-generated scores compared to human evaluations and explored how various prompting strategies influenced ChatGPT’s performance. Results revealed that ChatGPT could serve as an effective tool for L2 writing assessment when paired with carefully designed prompts. The study offers valuable insights into the strategic application of AI in educational settings.

Kim et al. (2024) examined the accuracy and reliability of ChatGPT's scoring compared to human ratings under two conditions: with and without writing prompts and source texts. Using a mixed methods approach, the study investigated rating discrepancies and their underlying causes. ChatGPT assessed 74 argumentative essays from the Iowa State University English Placement Test Corpus of Learner Writing under different prompting conditions, with its scores compared to those of human raters. The findings revealed that ChatGPT demonstrated moderate to low reliability across both conditions. Additionally, a qualitative analysis of its scoring rationales highlighted ChatGPT’s challenges in detecting content-related issues and effectively incorporating information from source texts, unlike human raters. The study

concluded that more rigorous training is required to align ChatGPT's evaluation processes with those of human raters.

Pfau et al. (2023) investigated the error detection capabilities of ChatGPT 3.5 Turbo compared to human raters using a corpus of essays written by Greek English learners across various proficiency levels. Their findings showed that, while ChatGPT occasionally missed errors, it exhibited a strong correlation with human raters (r = 0.97). They concluded that although human oversight is still necessary, ChatGPT significantly improves efficiency in identifying errors. Similarly, Jiang et al. (2023) evaluated ChatGPT alongside three other AI tools for error detection in L2 learners' texts. Their study revealed that AI models demonstrated high accuracy, with most achieving an accuracy rate of approximately 0.8.

In a separate study on the use of ChatGPT as an assessment tool for English language learners, Mizumoto and Eguchi (2023) applied an IELTS TASK 2 rubric as a prompt to evaluate 12,100 essays from the TOEFL11 dataset. These essays, previously rated by human evaluators, were classified into low, medium, or high proficiency levels on a five-point scale (Blanchard et al., 2013), though details about the scoring process were limited. The findings indicated that while ChatGPT achieved acceptable reliability, additional statistical adjustments were required to align its scores with a Quadratic Weighted Kappa (QWK) metric. This highlights the technical expertise needed to ensure accuracy and reliability, reducing the perceived advantage of using tools like ChatGPT for automated scoring. Concerning the shortcomings of the mentioned research, and in line with the aims of the study, the following research question (RQ) was posed:

RQ: Is there a significant correlation between IELTS essay scores scored by human raters compared with ChatGPT when tasks are sequenced differently and given corrective feedback?

Method

Research Design

The present research employed a correlational design to investigate the correlation between human ratings in comparison with ChatGPT in scoring IELTS when tasks were sequenced in different orders, and corrective feedback was also delivered to the participants. Human rating and ChatGPT rating were the independent variables, and IELTS scoring was the dependent variable. Task sequencing and corrective feedback were moderator variables. It is worth mentioning that this article is part of a larger project.

Participants

The study involved 113 sophomores who studied Teaching English as a Foreign Language at a private university in Isfahan, Iran. They enrolled in the Essay Writing course for the winter semester of 2022. According to the schedule, they were assigned to two groups. They were in their early twenties, with an average of 21. They had never been to English-speaking countries. They were supposed to become English teachers and had experience attending teacher training courses. They had passed courses on general English (grammar, listening and speaking, reading comprehension, and general English) before the study. They convened weekly for 90 minutes during the writing course. Their homogeneity was verified by administering the Oxford Quick

Placement Test. The results of the test indicated that their proficiency levels fell within the A2-B1 range based on the Community of European Framework of Reference.

Materials and Instruments

Academic Writing Coursebook

The Longman Academic Writing Series 3 (Hogue, 2013) was utilized as the primary textbook for the current study. The initial section of this textbook presents guidance on how to craft different kinds of academic paragraphs. It covers various types of paragraphs, such as process, definition, narrative, cause/effect, and comparison/contrast paragraphs. The subsequent section discusses fundamental concepts related to essay writing. The final section includes appendices and an index, making the book a convenient reference resource.

Writing Tasks

The syllabus for the course was designed using Longman Academic Writing Series 4 as its foundation. It included rules for punctuation, transitions, coherence and cohesion, the structure of essays and paragraphs, different types of essays, as well as techniques for paraphrasing and summarizing. Additionally, grammatical topics such as dangling modifiers, subject-verb agreement, reduction of relative clauses, and the use of prepositions were practiced through exercises.

In accordance with the IELTS writing task and following Malicka (2020), two groups of IELTS writing tasks were created. Each group included three tasks of varying degrees of difficulty. The levels of difficulty were modified based on the SSARC model. To achieve this, both the resource dispersing (R-dis) and resource directing (R-dir) variables from Robinson's TCF (2001) were taken into account while creating the tasks. Consequently, the decision to allow planning time was related to the R-dis variable, while the number of elements, specifically pictures in this instance, pertained to the R-dir variable.

Participants were required to articulate the tourist sites depicted in the images for the initial set of tasks. The following set of tasks presented images of routine activities and asked them to describe how these activities would be altered during their travels. For instance, one image focused on housing and lodging. In this image, participants were expected to articulate how their accommodations during trips differ from their daily lives.

In the initial stage of the model, known as stabilize simplify (SS), tasks were designed to be straightforward concerning planning time (R-dis). Thus, participants were allocated a planning period. Additionally, only two images were shown to the participants to maintain the task's simplicity in terms of the number of elements (R-dir). In the second stage, automatization (A), the second task of every set became more intricate regarding R-dis elements; therefore, no planning time was allowed for these tasks, although the number of elements remained unchanged. In the final stage of the model, reconstruct complexify (RC), the images were increased to four in order to raise the number of elements (R-dir), and no planning time was imposed for the tasks to ensure they remained complex concerning R-dis elements.

Procedure

Participants took a pretest at the beginning of the study. The pretest asked them to write about their favorite holiday. They had 28 minutes to plan, draft, and write the text. The researcher took and collected the tests for later analysis.

In the treatment phase, each group was given the tasks in a distinct order (S-C or C-S). Pictures were exposed to them by a video projector. The task requirement was explained, and A4 pieces of paper were given to them by the first researcher. Timing, for both planning and the main task, was considered, and they were asked to raise their paper when the time was over. The researcher watched them as an invigilator and ensured they did not talk to each other.

The researcher carefully reviewed the written texts and provided corrective feedback before returning the commented tasks to participants in the subsequent session. Within a constrained timeframe, participants were instructed to revise errors by writing corrections, leaving any unclear or unrecognized mistakes unchanged. Following a one-week interval, learners completed a comparable posttest under identical time constraints and conditions as the pretest, though with a distinct prompt asking them to describe their favorite travel destination. After collection, the researcher systematically prepared all responses for scoring. To ensure accuracy in textual analysis, she utilized Google Lens for digitization while meticulously verifying each transcribed version against the original handwritten submissions. After that, the participants' pre-tests and post-tests were blinded by the researcher and delivered to her colleagues. They were two English teachers with about five years of experience. They taught IELTS preparation courses and were familiar with IELTS band descriptors. They read and scored the participants' pretest and posttest based on the IELTS writing band descriptors. As a result, two scores were considered for each participant by each of the raters.

After the scores were obtained, and to ensure inter-rater reliability, a third teacher determined the final score for each participant; accordingly, two scores were finally considered as each participant's pre-test and post-test.

Apart from the teachers, ChatGPT version 4 rated the samples. In so doing, ChatGPT's settings were customized by providing details about the situation. In response to the first question, the researcher would like ChatGPT to know about her, information about the research goals, the researcher's job as an IELTS instructor, and her focus on IELTS writing band descriptors, as the assessment criteria were supplied. Similarly, regarding the question of how she would like ChatGPT to respond, information about the favorable qualities of the response was provided.

After that, the researcher made sure that Chat GPT was aware of the context by asking the bot general and detailed questions about it, and checking if its answers were right. Then, the prompt" Be a professional IELTS examiner. Read the texts and score them based on the IELTS writing band descriptors" was given to the bot. Later on, the sample texts for each participant were copied into the chat box, and each time, the same prompt was repeated after each set of texts. At times, when the bot ran into issues or seemed to generate irrelevant answers, the researcher paused and resumed this task after a while. Finally, the produced responses were scrutinized and, in case irrelevant, regenerated. The final responses were copied into a Docx

file. Two scores were calculated for each participant's pretest and posttest, respectively, by the robot rater. After gaining scores from the assessors and the bot, the researcher kept the numeric data for running statistical tests.

Results

This study investigated the correlation between human and ChatGPT scoring. As mentioned earlier, two sets of tasks were presented to the participants, with one set from simple to complex and the other set from complex to simple. The task requirement was explained. Later, the researcher read the texts and inserted corrective feedback on writing tasks. One week later, the participants took a posttest. The pretest and posttests were both scored based on the ILTS writing rubrics from band 1 to band 9. In order to investigate if there was any significant relationship between these variables, a *Pearson Correlation* test was carried out. The results obtained from the data are presented in this section.

Table 1. *Correlation between Human and ChatGPT IELTS Writing Scoring in S-C Tasks*

		Human_Scoring	**ChatGPT_Scoring**
Human_Scoring	Pearson Correlation	1	.968**
	Sig. (2-tailed)		.000
	N	113	113
ChatGPT _ Scoring	Pearson Correlation	.968**	1
	Sig. (2-tailed)	.000	
	N	113	113
**. Correlation is significant at the 0.01 level (2-tailed).			

Table 1 presents the results of a Pearson correlation between the scores from ChatGPT and those from human raters when tasks were presented from simple-to-complex along with corrective feedback from teachers. As shown in Table 1, there is a significant positive correlation (r =968, p > 05) between IELTS scores when human raters rate IELTS writing in comparison with the time when ChatGPT scores them and when tasks are presented from simple to complex. Figure 1 presents the scattergram of the correlation.

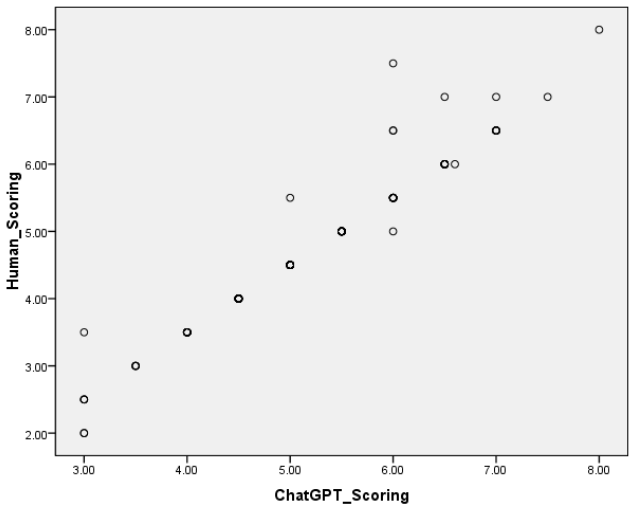

Figure 1. *Descriptive Statistics of the Pre-test Scores*

Figure 1 shows that the amount of *r* was found to be near one; therefore, all the points fell near the line of best fit. In order to obtain more reliable results, in the second phase of statistical analysis, the correlation coefficient between the IELTS writing scores, when tasks were presented from complex to simple, as rated by human raters and ChatGPT, was calculated. Table 2 presents the results.

Table 2. *Correlation between Human and ChatGPT IELTS Writing Scoring in C-S Tasks*

		Human _ Scoring CS	**ChatGPT _ Scoring CS**
Human _Scoring CS	Pearson Correlation	1	.860**
	Sig. (2-tailed)		.000
	N	113	113
ChatGPT _Scoring CS	Pearson Correlation	.860**	1
	Sig. (2-tailed)	.000	
	N	113	113
**. Correlation is significant at the 0.01 level (2-tailed).			

As shown in Table 2, a significant positive correlation (r = 860, p > 05) was observed between IELTS writing scores when rated by human raters in comparison with when they were scored by ChatGPT. Figure 2 presents the scattergram of the correlation coefficient.

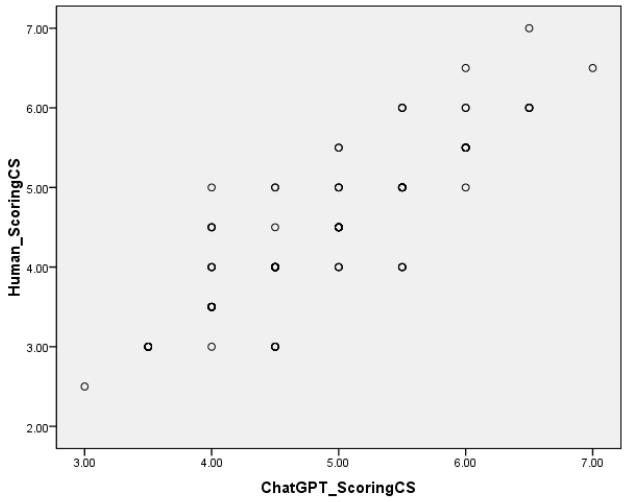

Figure 2. *Correlation between Human and ChatGPT IELTS Writing Scoring in C-S Tasks*

Figure 2 shows that in comparison with simple to complex task sequences, the amount of *r* was smaller. Accordingly, the points were scattered farther from the line of best fit.

Discussion

The present research intended to investigate the correlation between human and ChatGPT scores when tasks are presented from simple to complex or vice versa, along with corrective feedback. The findings indicated agreement between the scores presented by human raters and the bot. The findings about the applicability of ChatGPT can be justified in light of several theories. First, Sociocultural Theory highlights the contribution of tools and mediation in the

process of teaching and learning. ChatGPT has a mediatory role in the evaluation process by providing learners with unbiased feedback, which is in line with the principles of sociocultural theory. Moreover, ChatGPT's ability to provide consistent scoring without fatigue supports the reliability of assessments. It's programming to align with specific rubrics or criteria can also enhance validity. The next theory that supports the results of the present research is the Natural Language Processing (NLP) and AI Theories. The underlying theories of NLP and AI, such as machine learning and neural networks, provide the foundation for ChatGPT's ability to analyze and score writing tasks effectively.

Moreover, the results lend support to the study by Uchida (2024), which investigates the extent to which ChatGPT can accurately assess learners' writing and speaking. The researchers analyzed a total of 140 instances of writing and speaking data using ChatGPT and rated 80 by human evaluators. The results revealed the highest correlation between the overall writing scores and the scores given by ChatGPT.

The findings are comparable with several previous studies, such as Naismith et al. (2023), which reported a strong agreement between the two sets of scores. Similarly, the results support Jiang et al. (2023), who used ChatGPT alongside three other AI tools to identify errors in L2 Chinese writers' texts. Their research demonstrated high accuracy among AI models and a strong correlation between AI-generated and human-generated ratings. It is important to note that there were differences in their study and ours: samples were gathered online. Besides, the measures used to assess writing were more technical. Nevertheless, their findings were backed by our research.

In contrast, the findings of the present research contradict those of Kim et al. (2024), who examined the accuracy and reliability of ChatGPT-generated scores compared to human ratings. Their study reported that ChatGPT exhibited moderate to low reliability and was confined to identifying content-related issues and combining information from source texts. This discrepancy may stem from differences in the design of the two studies, as well as variations in the prompting conditions employed. These factors likely contributed to the contrasting outcomes.

Similarly, the results of the present study are inconsistent with the study by Shermis (2024), which aimed to determine if ChatGPT could match the scoring accuracy of human and machine scores. ChatGPT's performance was evaluated against human raters using quadratic weighted kappa (QWK) metrics. Results indicated that while ChatGPT's gradient boost model achieved QWKs close to human raters for some data sets, its overall performance was inconsistent and often lower than human scores.

The findings of the present research are also in contradiction with the study by Mizumoto and Eguchi (2023). The researchers applied an IELTS TASK 2 rubric as a prompt to evaluate 12,100 essays from the TOEFL11 dataset. These essays, previously rated by human evaluators, were classified into low, medium, or high proficiency levels on a five-point scale (Blanchard et al., 2013). The findings indicated that while ChatGPT achieved acceptable reliability, additional statistical adjustments were required to align its scores with a Quadratic Weighted Kappa (QWK) metric. This highlights the technical expertise needed to ensure accuracy and

reliability, reducing the perceived advantage of using tools like ChatGPT for automated scoring.

In addition, there is a contradiction between the results of the present study and the study by Bui and Barrot (2024), which investigates the correlation between scores assigned by ChatGPT and those given by a human evaluator, as well as the consistency of ChatGPT's scores across multiple assessments. A cross-sectional quantitative methodology was utilized to analyze a total of 200 argumentative essays, with 50 essays selected from each proficiency level (A2_0, B1_1, B1_2, and B2_0). Both ChatGPT and a seasoned human rater evaluated these essays. The findings, derived from correlational analysis, indicated that the scoring by ChatGPT did not closely correspond with that of the experienced human rater, demonstrating weak to moderate relationships. Furthermore, the analysis revealed a lack of consistency in ChatGPT's scores after two rounds of evaluation, as evidenced by low intraclass correlation coefficient values.

Ramineni and Williamson (2018) argued that using ChatGPT as a second coder may help identify potential errors or biases for classroom-based assessments. However, the results of the present research revealed that besides using ChatGPT as a second coder, it could be used as a self-assessment instrument for language learners. This finding could be a complement to the previous one, as it is generally in line with that.

Conclusion

The findings revealed that ChatGPT can score essays written by human raters almost accurately. However, it should be noted that human editing may still be needed in some cases. Educational Testing Service (ETS) and other testing organizations frequently contend that AI instruments should be utilized solely as second coders (Ramineni & Williamson, 2018).

As in modern life, teachers may rarely have time to assess classroom tasks, ChatGPT could also be used as a writing assessment tool to save time. According to existing research, writing in a second language can help language learning (Polio & Park, 2016). On the other hand, teachers are unwilling to assign writing without assessments. This is the point at which AI tools such as ChatGPT should enter. L2 learners can take advantage of the self-assessment tools and assess their writing using ChatGPT. According to Poole and Polio (2024), this approach not only increases the volume of writing that learners engage in but also promotes the development of metacognitive skills and digital literacy related to the use of new AI tools.

Besides using ChatGPT as a second coder, it can be employed as a self-assessment instrument for language learners. Some studies (e.g., Polio & Park, 2016) have indicated that writing in an L2 can help L2 learners. Language teachers may lack the time to evaluate class writing essays. In this context, using ChatGPT for self-assessment may not only enhance the amount of writing that learners produce but also support the development of metacognitive and digital literacy skills (Poole & Polio, 2024).

Similar to other studies, the present study suffers from several limitations. First, there was a limited range of scores, and only 113 participants. Second, the participants were intermediate-level learners. Moreover, in the present research, only one type of AI tool, namely, ChatGPT, was investigated. Other tools can be the subject of future research.

The findings from both human ratings and ChatGPT ratings indicate the potential of employing GPT for automatic scoring in instructional settings. Pre-trained models can be effectively tailored to specific domain tasks through the optimization technique known as fine-tuning. The results of this study indicate that AI-based scoring can enhance accuracy in writing tasks. Subsequent research could investigate the application of automatic scoring in real classroom environments and evaluate student learning outcomes through experimental design.

As enthusiasm grows around AI tools like ChatGPT for writing assessments, educators must understand both their potential and their limitations. Familiarity with the strengths and weaknesses of ChatGPT, especially in evaluating student writing, enables educators to make well-informed choices about incorporating these tools into their teaching and assessment practices. Recognizing how AI can contribute, such as in spotting grammatical errors, is crucial for effective integration.

Acknowledgments (Required)

The authors would like to express their heartfelt appreciation to the contributors and participants whose assistance was vital to this research. They are also grateful to the reviewers and editorial board for their valuable contributions and feedback on the manuscript.

References

Allaw, E., & McDonough, K. (2019). The effect of task sequencing on second language written lexical complexity, accuracy, and fluency. *System*, 85(2019), 102-104. https://doi.org/10.1016/j.system.2019.06.008

Bui, N. M., & Barrot, J. S. (2024). ChatGPT as an automated essay scoring tool in the writing classrooms: How it compares with human scoring. *Education and Information Technologies,* 30(2041–2058). https://doi.org/10.1007/s10639-024-12891-w

Cotos, E. (2014). Genre-based automated writing evaluation for L2 research writing. Palgrave Macmillan. Dai, W., Lin, J., Jin, H., Li, T., Tsai, Y. S., Gašević, D., & Chen, G. (2023). Can large language models provide feedback to students? A case study on ChatGPT. In 2023 *IEEE International Conference on Advanced Learning Technologies* (ICALT),323-325.

Kim, H., Baghestani, Sh., Yin, Sh., Karatay, Y., Kurt, S., Beck, J., & Karatay, L. (2024). ChatGPT for writing evaluation: Examining the accuracy and reliability of AI-generated scores compared to human raters. In C. A. Chapelle, G. H. Beckett, & J. Ranalli (Eds.), *Exploring artificial intelligence in applied linguistics* (pp. 73-95). Iowa State University Digital Press. https://doi.org/10.31274/isudp.2024.154.06

Kumaravadivelu, B. (2006b). TESOL methods: Changing tracks, challenging trends. *TESOL Quarterly*, 40(1), 59- 81. https://doi.org/10.2307/40264511

Lagakis, P., & Demetriadis, S. (2021). Automated essay feedback generation in the learning of writing: A Review of the Field. *Interactive Mobile Communication, Technologies and Learning*, 2(1), 443-453. https://doi.org/10.1007/978-3-030-96296-8_40

Liu, Q., & Brown, D. (2015). Methodological synthesis of research on the effectiveness of corrective feedback in L2 writing. *Journal of Second Language Writing*, 30(2015), 66-81. https://doi.org/10.1016/j.jslw2015.08.011

Lu, X., & Hu, R. (2022). Sense-aware lexical sophistication indices and their relationship to second language writing quality. *Behavior research methods*, 54(3), 1444-1460. https://doi.org/10.3758/s13428-021-01675-6

Nam, B. H., & Bai, Q. (2023). ChatGPT and its ethical implications for STEM research and higher education: A media discourse analysis. *International Journal of STEM Education*, 10(66). https://doi.org/10.1186/s40594-023-00452-5

Baralt M., Gilabert R., Robinson P. (2014). An introduction to theory and research in task sequencing and instructed second language learning. In Baralt M., Gilabert R., Robinson P. (Eds.), *Task sequencing and instructed second language learning* (pp. 1–37). Bloomsbury.

Blanchard, D., Tetreault, J., Higgins, D., Cahill, A., & Chodorow, M. (2013). TOEFL11: A corpus of non-native English. *ETS Research Report Series* (i-15). https://doi.org/10.1002/j.2333-8504.2013.tb02331.x

Hogue, C., Fry, M. & Fry, A. & Pressman, S. (2013). The Influence of a Motivational Climate Intervention on Participants' Salivary Cortisol and Psychological Responses. *Journal of sport & exercise psychology*. 35*(1).* 85-97. 10.1123/jsep.35.1.85.

Jiang, Z., Xu, Z., Pan, Z., He, J., & Xie, K. (2023). Exploring the role of artificial intelligence in facilitating assessment of writing performance in second language learning. *Languages*, 8(4), 247-264. https://doi.org/10.3390/languages8040247

Malicka, A. (2020). The role of task sequencing in fluency, accuracy, and complexity: Investigating the SSARC model of pedagogic task sequencing. *Language Teaching Research*, 24(5), 642-665. https://doi.org/10.1177/1362168818813668

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2), 23-41. https://doi.org/10.1016/j.resmal.2023.100050

Naismith, B. & Mulcaire, P. & Burstein, J. (2023). *Automated evaluation of written discourse coherence using GPT-4*. Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), 394-403. 10.18653/v1/2023.bea-1.32.

Nguyen, N. H., & Nguyen, K. D. (2022). Vietnamese learners' performance in The IELTS writing task 2 *International Journal of TESOL & Education*. 2(1)170-189. https://doi.org/10.54855/ijte.222111

Oshima, A., & Hogue, A. (2013). *Longman Academic Writing Series, Level 3: Paragraphs to Essays*. (3rd ed.). Longman

Pfau, A., Polio, C., & Xu, Y. (2023). Exploring the potential of ChatGPT in assessing L2 writing accuracy for research purposes. *Research Methods in Applied Linguistics*, 2(3), 45-67. https://doi.org/10.1016/j.resmal.2023.100083

Polio, C., & Park, J. H. (2016). Language development in second language writing. In Manchón, R. M. & Matsuda, P. (Eds.). *Handbook of Second and Foreign Language Writing*. de Gruyter, pp. 287–306. https://doi.org/10.1515/9781614511335-017

Poole, F. M, & Coss, M. D. (2024). Can ChatGPT Reliably and Accurately Apply a Rubric to L2 Writing Assessments? The Devil is in the Prompt(s). *Journal of Technology and Chinese Language Teaching*, 15(1), 1-24. http://www.tclt.us/journal/2024v15n1/poolecoss.pdf

Poole, F. J., & Polio, C. (2024). From sci-fi to the classroom: Implications of AI in task-based writing. *TASK: Journal on Task-Based Language Teaching,* 3(2), 243-272. https://doi.org/10.1007/s10462-021-10068-2

Ramineni, C., & Williamson, D. (2018). Understanding mean score differences between the e-rater® automated scoring engine and humans for demographically based groups in the GRE® general test. *ETS Research Report Series,* 2018(1), 1–31. https://doi.org/10.1002/ets2.12211

Rattanadilok, P. (2015). Understanding EFL students' errors in writing. *Journal of Education and Practice,* 6(32), 99-106. https://files.eric.ed.gov/fulltext/EJ1083531.pdf

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied linguistics,* 22(1), 27-57. https://doi.org/10.1093/applin /22.1.27

Robinson, P. (2007). Task complexity, theory of mind, and intentional reasoning: Effects on L2 speech production, interaction, uptake and perceptions of task difficulty. *International Review of Applied Linguistics in Language Teaching,* 45, 193-213. https://doi.org/10.1515/iral.2007.009

Shermis, M. D. (2024). *Using ChatGPT to score essays and short-form constructed responses*. ArXiv. https://doi.org/10.48550/arXiv.2408.09540

Tang, C., & Liu, Y. T. (2018). Effects of indirect coded corrective feedback with and without short affective teacher comments on L2 writing performance, learner uptake and motivation. *Assessing Writing*, 35, 26-40. https://doi.org/10.1016/j.asw.2017.12.002

Tajeddin, Z., & Bahador, H. (2012). Pair grouping and resource-dispersing variables of cognitive task complexity: Effects on L2 output. *Iranian Journal of Applied Linguistics (IJAL)*, 15(1), 123-149. http://ijal.khu.ac.ir/article-1-81-en.html

Uchida, S. (2024). Evaluating the accuracy of ChatGPT in assessing writing and speaking: A verification study using ICNALE GRA. *Learner Corpus Studies in Asia and the World,* 6(1), 1–12. https://doi.org/10.24546/0100487710

Zhai, K., & Gao, X. (2018). Effects of corrective feedback on EFL speaking task complexity in China's university classroom. *Cogent Education*, 5(1), 148-157. https://doi.org/10.1080/2331186X.2018.1485472