International Journal of Assessment Tools in Education
2025, Vol. 12, No. 1, 62–77
https://doi.org/10.21449/ijate.1496193
journal homepage: https://dergipark.org.tr/en/pub/ijate
Research Article

ChatGPT-3.5 as an automatic scoring system and feedback provider in IELTS exams

Xinming Chen[1*], **Ziqian Zhou**[2], **Malila Prado**[3]

[1]University College London, England
[2]The University of Hong Kong, Hong Kong (SAR) China
[3]BNU-HKBU United International College, China

ARTICLE HISTORY
Received: June 6, 2024
Accepted: Nov. 17, 2024

Keywords:
ChatGPT-3.5,
Automatic Essay Scoring,
AI Proofreader,
IELTS,
L2 learning.

Abstract: This study explores the efficacy of ChatGPT-3.5, an AI chatbot, used as an Automatic Essay Scoring (AES) system and feedback provider for IELTS essay preparation. It investigates the alignment between scores given by ChatGPT-3.5 and those assigned by official IELTS examiners to establish its reliability as an AES. It also identifies the strategies employed by ChatGPT-3.5 in revising essays based on the four IELTS rubrics: task achievement, coherence and cohesion, lexical resources, and grammatical range and accuracy. Based on pre-rated essays from an official IELTS preparatory book as a control measure to ensure objectivity, the findings indicate a discrepancy, with ChatGPT-3.5 typically assigning lower scores compared to official raters. However, ChatGPT-3.5 shows a robust capability to revise essays across all four descriptors. In addition, the effectiveness of ChatGPT-3.5 as a feedback provider may be attributed to the essay type and its widely accepted rubrics. Our study contributes to the understanding of the application of AI tools in second language writing and suggests that future studies should focus on evaluating the capacity and effectiveness of such tools in pedagogical applications.

1. INTRODUCTION

Even though technology is becoming increasingly present in education, it does not appear to have dramatically changed the way we teach. Though technology is at every teacher's disposal, old pedagogical concepts appear to meet the needs of most teachers (Chiu *et al.*, 2023). Regarding English as a Second Language (ESL) writing, there is resistance among teachers against employing technology such as Grammarly (Huang *et al.*, 2020), machine translation (Lee, 2023), or even using digitally available multilingual resources (Prado & Huggins, 2023). Chiu *et al.* (2023) report that "some teachers described the technologies as difficult to control, lacked an understanding of how the technologies operated, and were concerned about ethical issues, such as bias and breaches of privacy." This probably explains why the response to the launch of ChatGPT (Open AI, 2022) at the end of 2022 was not widely embraced in the education realm, particularly in higher education.

*CONTACT: Xinming CHEN q030025010@mail.uic.edu.cn University College London, Faculty of Education and Society, Department of Psychology and Human Development, England

e-ISSN: 2148-7456

As suggested by Bai *et al*. (2022), Artificial Intelligence (AI) applications in Education (AIEd) are a trending research topic. ChatGPT, an artificial intelligence chatbot, uses natural language processing to create humanlike conversations based on large amounts of digital content (Boa Sorte *et al.*, 2021; Pavlik, 2023; Fryer *et al.*, 2020). It can compose texts in a variety of written genres, including articles, social media posts, essays, and emails, all generated in a conversation-like style (Boa Sorte *et al.*, 2021). However, the introduction of ChatGPT in academia has sparked debates regarding authorship and concerns over plagiarism (Dergaa *et al.*, 2023) and raised the concern that teachers might be substituted (Warschauer *et al.*, 2023).

Yet ChatGPT is having a significant impact on language education research, particularly in second language (L2) writing (Artiles Rodríguez *et al.*, 2021; Barrot, 2023; Baskara, 2023; Dergaa *et al.*, 2023; Han *et al.*, 2023; Warschauer *et al.*, 2023). Four major advantages of ChatGPT as a writing assistant tool have been considered: i) providing instant and realistic interactions with learners; ii) designing personalized learning materials based on different proficiency levels; iii) stimulating learners' interests; and iv) providing timely and adaptive feedback and assessments (Barrot, 2023; Fryer *et al.*, 2020; Huang *et al.*, 2022; Kuhail *et al.*, 2023). While ChatGPT has been shown to be a productive tool for students whose English is not their first language (L1), a few scholars have argued against it because it will either cut down on the practice of good writing demands or hinder creative or critical thinking skills (Liang *et al.*, 2023).

The workload of writing classes for teachers consists of a large amount of assessment, including review, feedback, and grading. In large classes, the task becomes impractical. A solution to this problem may be the use of AI technology such as ChatGPT (Kohnke *et al.*, 2023), which enables the provision of autonomous feedback to students (Artiles Rodríguez *et al.*, 2021; Ranalli, 2018). However, reducing the teacher's workload through automated marking or teaching students to grade themselves poses several challenges, including issues of reliability, consistency, and quality. While educational and linguistic software packages are available for automated assessment and grading, such as Pigaiwang and Coh-Metrix (Zhou & Prado, 2024), the functionality of chatbots allows for easier consultation between the student and the tool and, as such, more effective use of these tools, thus aiding in the management of assessments. In response, we suggest that using chatbots can significantly simplify the task of grading, thereby lessening teachers' workload.

This study explores the use of ChatGPT-3.5 as automated feedback on writing system (Cotos, 2023) and a proofreader for assessing and revising students' essays. In pursuit of objectivity and reliability in our analysis, this study makes use of essays sourced from an official preparatory book for the International English Language Testing System (IELTS), one of the world's most widely accepted English proficiency exams. These essays, previously assessed and selected by IELTS examiners for publication, served as a benchmark for evaluating ChatGPT-3.5's scoring reliability. The choice to use pre-rated essays aims to mitigate the potential subjectivity associated with individual rater judgments. By relying on essays with established scores, we created a more controlled environment to investigate the consistency and reliability of ChatGPT-3.5 as a scoring mechanism as against the standardized criteria set by IELTS, whose descriptors (task achievement, coherence and cohesion, lexical resources, and grammatical range and accuracy) are already embedded in ChatGPT. This methodological approach ensures that the evaluation of ChatGPT-3.5's effectiveness as an Automated Essay Scoring (AES) system is grounded in comparison with authoritative, pre-validated assessments, thus providing a foundation for our analysis. The study manually and qualitatively classifies the strategies used by ChatGPT-3.5 in revising examinees' essays in terms of the four descriptors in the IELTS rubrics, identifying the strengths and weaknesses of ChatGPT-3.5 for revising the essays against different descriptors. To this end, the study investigated the following research questions:

- To what extent do scores on essays differ (or are consistent) between ChatGPT-3.5 and official raters?
- What strategies are used by ChatGPT-3.5 to revise students' IELTS essays?

The results of this study will serve to advance educators' awareness of the pros and cons of ChatGPT as an AES and proofreader. Furthermore, the study will provide directions for future research in the application of ChatGPT to L2 writing. The findings will also shed light on the pedagogical implications of the use of AI tools in future education.

2. LITERATURE REVIEW

2.1. Automatic Scoring and Evaluation of Writing

Traditional classroom-based teaching of writing lacks individual attention to students' learning, resulting in a lack of autonomy and self-initiative, with students passively waiting for teachers to assign essays to be later graded (Yang & Dai, 2015). Automated Essay Scoring (AES) refers to the use of specialized computer programs to evaluate and score the characteristics of compositions based on validity, impartiality, and reliability (Shermis & Burstein, 2003). The development of such systems is the embodiment of the development of machine-assisted language testing (Yang & Dai, 2015), which, technically, is usually based on mathematical formulas and equations for linguistic decodings of the textual features (Zhou & Prado, 2024). In the 1960s, the development of Page Essay Grade (PEG), a program that used multiple regression analysis of measurable text features to build a scoring model based on a corpus of essays previously graded by hand, marked the beginning of AES (He, 2016; Mizumoto & Eguchi, 2023). A large number of AES programs, such as Criterion, My Access, Writing Roadmap, and Pigaiwang, followed suit. These programs were equipped with a number of functions, including a scoring engine, an editing tool that offered grammar and spelling feedback, and a dictionary (He, 2016). As proposed by Bai *et al*. (2022), AES systems are able to lower teachers' workload, especially in situations where learning needs are highly specific.

He (2016) classified the research in AES systems into three types: i) validity of the software; ii) learning outcomes and improvements to learners' writing skills; and iii) use of writing software tools in classroom settings. One of the most recent research projects, carried out by Mizumoto and Eguchi (2023), was representative of the first type. They collected 12,100 Test of English as a Foreign Language (TOEFL) essays and compared the scores given by ChatGPT-3.0 with the benchmark levels, aiming to explore the reliability and accuracy of using ChatGPT-3.0 as an AES along with the linguistic features that influenced the system itself. Their results showed that ChatGPT had a certain level of accuracy and reliability. Moreover, Mizumoto and Eguchi considered several linguistic features at the level of lexis, phraseology, syntax, and cohesion based on previous research that investigated linguistic correlates of human rating scores. They found that the more linguistic features of a text were taken into consideration while evaluating, the more accurate this was reflected in the scoring.

Studies of the second type, namely research in students' learning outcomes, are exemplified by the longitudinal research carried out by Huynh-Cam *et al.* (2023) on students' writing quality. These researchers collected the English writing scores of 82 university students before and after the intervention of an AES tool named Marking Mate in a course of English as a Foreign Language (EFL) writing. A self-report survey was also conducted to explore the attitude of students toward studying with this AES tool. The study found a rise in writing scores using the AES tool as well as favorable opinions from students toward the usefulness of the tool. As regards the third type of AES research, namely its implementation in the classroom, Li (2021) investigated how teachers perceived ESL writing classes supported by Criterion, an automated writing evaluation system. The research found that different teachers tended to take different approaches to implementing the same evaluation tool in classrooms, which in turn reflected observable differences in writing quality. This advocates for the value and significance of teacher agency and cognition in the AES-assisted English teaching classroom.

Having derived from AES, Automated Writing Evaluation (AWE) tools "support the process of writing by providing formative feedback that is typically displayed on an engaging graphic interface" (Cotos, 2023, pp. 347–348). Such tools, considered formative while AES tools are summative (Cotos, 2023), go several steps further in that they employ AI to generate feedback on lexical, semantic, syntactic, and discourse elements on students' writing. AWE tools allow students to draft a text as many times as they wish and be agentive in their selection of feedback, which can vary from global writing skills to language mechanics (Stevenson & Phakiti, 2014).

However, the capabilities offered by AWE tools may not be easily accessed by students. In his L2 writing qualitative study of three students engaging with AWE feedback on their own writing, Zhang (2020) observed that even with a machine designed for the task of analyzing both micro- and macro-level issues, students had their attention drawn almost exclusively to micro-level changes such as spelling and grammar mistakes. In contrast, macro-level changes such as redundancy were attended to only once in Zhang's study, which may reflect a mutual correspondence with higher proficiency levels. Thus, according to Zhang, there is a need for a radical change in how we view L2 revision, which should diverge from an error-reduction activity in favor of more global development.

As regards the field of AIEd, Chiu *et al.* (2023) list several critical areas, among which is the implementation of AI technologies for automating student assessment and predicting their performance. According to their study, priority should be given to the development of a new pedagogical framework centered on AI learning and teaching, particularly in supporting teachers' assessment by "providing automatic marking and predicting students' performance" (p. 9) along with the application of personalized learning. Conditional on this objective is the importance of teachers themselves possessing sufficient knowledge of AI tools and their pedagogical applications. To this end, the authors suggest that future studies should concentrate on the evaluation of the capacity and effectiveness of AI tools applicable to pedagogy.

2.2. Chatbots to Support Writing Feedback and Improvement

Bašić *et al.* (2023) tested ChatGPT-3 as essay-writing assistance for students. The authors compared 18 second-year masters students' essay writing performance with or without employing ChatGPT-3 as a writing assistant tool. Results showed no evidence that using ChatGPT-3 improved the quality of students' essays. This result was consistent with the findings of Fyfe (2022), which tested students' use of GPT-2 and found that students regarded writing independently as easier than writing with GPT-2 as they would be distracted by the texts generated by GPT-2 for the writing task. The study concluded that the use of ChatGPT as an assistance tool could not reduce students' writing time. However, it is worth mentioning that in the study conducted by Bašić *et al.* (2023), the essays were written in Croatian rather than in English. Given that ChatGPT was predominantly fed with English content and thus may have generated higher-quality information in English for students who used it as an essay-writing assistant tool, the results may have been different if English essays had been used instead.

However, some studies support the view that ChatGPT may be beneficial to L2 writing. Han *et al.* (2023) investigated the integration of ChatGPT into L2 writing courses by creating a learning platform called RECIPE (Revising an Essay with ChatGPT) on an Interactive Platform with 213 EFL undergraduate and graduate learners. ChatGPT played the role of a personalized English writing teacher and instructed the students step by step on revising their writing. The results showed that this kind of learning could improve students' writing ability as the steps reminded students of the lecture content and helped them receive a more class-relevant response from ChatGPT. At the end of the course, students reflected that they had a positive experience working with ChatGPT.

Although the effectiveness of ChatGPT-2.0 or 3.0 in grading students' essays and being an assistant to students has been investigated, the quality and nature of improvements to reviewed

texts remain to be explored. It is important to examine the characteristics of the suggestions made by chatbots, such as ChatGPT, along with their reliability.

3. METHOD

ChatGPT-3.5, currently a free version, was employed to verify how consistent its suggestions are and to review the feedback it provides. To ensure data consistency, this study made use of one of the most widely used large-scale ESL tests with a writing test component, namely IELTS, the International English Language Testing System, a highly popular exam worldwide as well as in China. The writing section of IELTS contains two types of assignments. The first is a short essay that usually requires candidates to write about 150 words to describe data from a chart or table, and the second is an argumentative essay of about 250 words (for a critical review, see Uysal, 2010).

Bai *et al*. (2022) reviewed 13 studies of the assessing power and accuracy of AES tools in 2021 and found that different studies used different measures. They concluded that the simplest measures consist of focusing on the correlation between human and machine scoring (Pearson correlation coefficient R) and exact accuracy (i.e., the percentage of cases when both human and machine agree on the exact score). Following the same prompt, our study used a quantitative method that references the correlation between human IELTS examiners' grading and ChatGPT-3.5 scores to investigate any differences through experimental comparisons with Pearson's R. Furthermore, a qualitative method was also used focusing on the observation of the strategies used by ChatGPT-3.5 in revising the essays.

3.1. Resources

A total of 23 essays officially scored between band 5.5 and 6.5 were taken from Cambridge IELTS volumes 1 to 17 (see Table 1). The Cambridge IELTS consists of a selection of official examination papers from the University of Cambridge ESOL Examinations with the purpose of preparing candidates for the tests.

Table 1. *Selected essays from Cambridge IELTS Volumes 1-17.*

Publisher	Number	Volume	Year of First Publication	Test No.	Word Count	Score
Cambridge University Press & Cambridge English Language Assessment	1	3	2002	4	317	6
	2	3	2002	Training B	260	6
	3	4	2005	Training A	334	6
	4	5	2006	3	369	6
	5	6	2007	Training A	285	6
	6	8	2011	2	250	5.5
	7	8	2011	4	378	6.5
	8	9	2013	Training A	302	6
	9	10	2015	4	224	5.5
	10	11	2016	1	264	5.5
	11	11	2016	4	276	5.5
	12	12	2017	5	269	6
	13	13	2018	1	313	6.5
	14	13	2018	3	282	6
	15	13	2018	4	276	6
	16	14	2019	3	240	5.5
	17	15	2020	2	350	6
	18	15	2020	4	269	6.5
	19	16	2021	1	284	6
	20	17	2022	1	243	6.5
	21	17	2022	2	280	6.5
	22	17	2022	3	280	6.5
	23	17	2022	4	254	6

The texts were written by candidates and assessed by official IELTS examiners based on four descriptors: task achievement, coherence and cohesion, lexical resources, and grammatical range and accuracy. They are employed as examples or samples to be used by future candidates. These essays correspond to IELTS Writing Task 2, which aims to assess students' ability to provide solutions to problems, clearly presenting and justifying their opinions and supporting them with explicit, logical, and related evidence. Based on IELTS Test Demographic Data (Test Statistics, 2022),[†] which states that the largest proportion (62%) of IELTS scores received by candidates seeking a higher education course was between band 5.5 and 6.5, we selected scores ranging from bands 5.5 to 6.5.

3.2. IELTS Descriptors

As mentioned above**,** the IELTS writing exam consists of four descriptors: task achievement, coherence and cohesiveness, lexical resources, and grammatical range and accuracy. Grammatical range and accuracy are first and foremost a descriptor that emphasizes the accuracy and range of the grammar in the essay. For instance, candidates are expected to use complex structures, appropriate tenses, comparatives, conditionals, and modal verbs in their writing. Second, the lexical resources descriptor highlights the range and accuracy of vocabulary, including synonyms, collocations, and parts of speech. Coherence and cohesiveness, the third descriptor, refers to the flow of texts and how the paragraphs are structured. Finally, task achievement is concerned with how fully the exam question has been answered.

3.3. Instruments

To collect sufficient and useful data to answer the research questions, ChatGPT-3.5 and R were used as the instruments in this study. ChatGPT-3.5 was used to score the 23 essays and revise them to band 7. The suggestions generated by the chatbot were individually compared, and submitted to R for the descriptive data calculation. R is a computational language and a data processing, calculation, and mapping software system that is increasingly being used in research in many disciplines (Crawley, 2012). A further explanation of its use will be included in the next subsections.

3.4. Procedure

The research procedure was divided into two parts. The first part aimed to answer the first research question. After we collected a total of 23 sample essays with scores ranging from bands 5.5 to 6.5, we inserted them into ChatGPT for scoring.

The following steps were replicated with each of the 23 sample essays. First, we gave the chatbot a single prompt, consisting of the request, "Please give a score to this essay in terms of the four descriptors of IELTS writing rubrics", followed by each of the IELTS writing prompts and writing samples. The input is brief as we aimed to imitate how students or teachers, as real-life users, would make use of ChatGPT. For each essay, we input five times, and since, in some cases, the output results of the grade of the same essay were different, the average score of the grades provided in the five rounds was adopted as the grade for later data analysis. We then copied the average band score of each essay given by ChatGPT-3.5, and altogether, there were 23 scores given by ChatGPT-3.5. A t-test between the 23 official scores and the 23 ChatGPT-provided scores was performed through the R language software to ascertain whether there was a significant difference between the gradings of the two groups, namely the samples rated in the resource book and ChatGPT-3.5. In addition, we repeated these steps by inputting "Please give a score to this essay in terms of the four descriptors of IELTS writing rubrics" and the essay again, but this time, we did not provide GPT with the IELTS writing prompt, or the

[†] Text Statistics (2022): https://ielts.org/researchers/our-research/test-statistics#Demographic

required essay topic from the question. A paired t-test was performed again with this group of data and human ratings. This process helped us find whether GPT read and considered the required writing topic for grading.

The second part of the study addressed the second research question by analyzing the revision strategies adopted by ChatGPT. All the selected essays were inserted into ChatGPT-3.5 along with the new prompt "Please revise this IELTS essay to make it achieve a band score of 7 referring to the IELTS writing rubric." Subsequently, we selected 10 of the 23 revised essays through a systematic sampling method by publication year (Table 2), analyzed the revisions suggested by ChatGPT-3.5, coded and classified each revision in terms of the four descriptors from IELTS benchmark (task achievement, coherence and cohesion, lexical resources, and grammatical range and accuracy), with more detailed sub-categories under each descriptor. The analysis and classification achieved by the coding method were implemented through Microsoft Word, particularly the Highlight and Comment functions, to facilitate our collaborative analysis. We first conducted text analyses and coding independently for the ten essays, then discussed until we reached a baseline of 80% intercoder reliability, given that an 80% intercoder reliability is advocated as reliable by scholars such as Miles and Huberman (1994). A more detailed classification of the revisions was then made under each descriptor. Finally, we calculated the strategies most frequently used by ChatGPT-3.5 for further explanation.

As an additional step, despite sampling ten essays for further text analysis, we input all 23 essays into ChatGPT-3.5 for proofreading and revision, after which we input the revised essays again into ChatGPT-3.5 on a separate new page, asking it to assess and grade the revised essays. This helped us explore if the proofreading of ChatGPT-3.5 was effective from the view of ChatGPT-3.5 itself, as we would verify if there was a difference in grades between the original essays and the revised essays.

Table 2. *Selected essays for data analysis.*

Publisher	Number	Series	Year of First Publication	Test Number	Word Count	Score
Cambridge University Press & Cambridge English Language Assessment	1	3	2002	4	317	6
	2	4	2005	Training A	334	6
	3	6	2007	Training A	285	6
	4	8	2011	4	378	6.5
	5	10	2015	4	224	5.5
	6	11	2016	4	276	5.5
	7	15	2020	2	350	6
	8	16	2021	1	284	6
	9	17	2022	2	280	6.5
	10	17	2022	4	254	6

3.5 Data Analysis

We now outline the statistical methods used to analyze the data collected from the 23 IELTS essays, focusing on comparing the scores provided by ChatGPT-3.5 and official IELTS raters, as well as analyzing the revisions made by ChatGPT-3.5 in response to the essays.

The primary method for analyzing the scores given by ChatGPT-3.5 and official IELTS raters was the paired samples t-test, which was used to compare the scores of each essay between the two groups (ChatGPT-3.5 vs. IELTS official raters). The t-test helped us assess whether the differences between the two sets of scores were statistically significant. A paired t-test provides us with the gap between grades of every essay from the two groups rather than an overall distribution of scores of the two groups. This ensures that we focus on each essay in terms of the difference between the two raters, and the t-tests work as an investigator of the scoring gaps of all 23 essays rated by the two raters.

4. RESULTS

4.1. ChatGPT as an Automatic Essay Scoring (AES) System

The numerical data on the grading of the 23 essays are displayed in Table 3, which also shows the mean scores given by ChatGPT-3.5 (with the input of the required topic) and the Cambridge official examiners. Additionally, the table displays the *p*-value of students' t-tests comparing the scores given by ChatGPT-3.5 and those on the official resource book.

Table 3. *Mean scores and t-test (1).*

ChatGPT (input with topic)	Examiners	*p*-value
5.65	6	0.038

The *t*-test checked the degree of difference in the scores given by ChatGPT-3.5 (with input IELTS instructions) and those given by the official IELTS examiners. Results revealed a significant difference between the scores given by the two approaches: ChatGPT-3.5 (with instructions) (*M*=5.65, *SD*=0.93) and IELTS examiners (*M*=6.00, *SD*=0.34), *t*=-1.8606, *p*=.03843.

As mentioned earlier, to check whether ChatGPT-3.5 considered the instructions provided, a new round of testing was performed by inputting *without instructions* for each essay. The results of a *t*-test comparing the scores of ChatGPT-3.5 and those of IELTS examiners are shown in Table 4.

Table 4. *Mean scores and t-test (1).*

ChatGPT (input without instructions)	Examiners	*p*-value
5.75	6	0.077

The results also reveal a difference between the scores given by the two approaches with a 90% confidence interval. However, the difference between the scores provided by ChatGPT-3.5 (*M*=5.75, *SD*=0.86) and those of the examiners (*M*=6.00, *SD*=0.34) was smaller (*t*=-1.4735, *p*=.07772) compared with the difference shown in Table 4.

To test the difference in scores given by ChatGPT-3.5 with and without inputting instructions, a third t-test was performed by R. The two groups of grades are 1) GPT's grading with our input of the instructions from the writing question and 2) GPT's grading without our input of the writing instruction but only the request of grading and the sample essay.

Table 5. *t-test by R.*

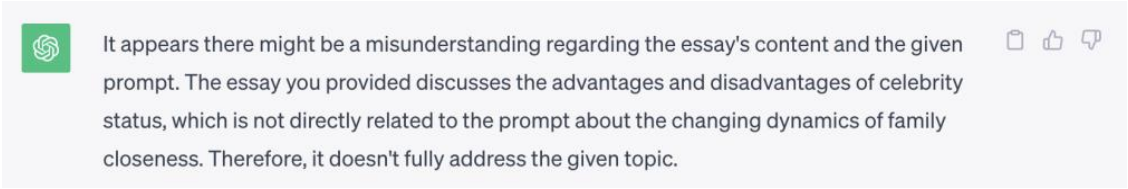
Group	Mean Score (*M*)	Standard Deviation (*SD*)	*t*-value	Degree of freedom	*p*-value
ChatGPT-3.5 (with instructions)	5.65	0.93	-1.0058	21	0.326
ChatGPT-3.5 (without instructions)	5.75	0.86			

The results show no significant difference between the scores given by the two approaches. To be specific, there was no evidence of a significant difference between the scores provided by GPT-3.5 *without* instructions (*M*=5.75, *SD*=0.86) and *with* instructions (*M*=5.65, *SD*=0.93) over short-term learning outcomes (*t*=-1.0058, *p*=.326). This indicates that whether inputting the required writing topic or not, GPT will grade the essay similarly, with almost the same scores.

During the interaction with ChatGPT-3.5, two responses were noted. First, even though there was no significant difference between providing and not providing instructions for the essays,

this does not imply that ChatGPT-3.5 disregards the instructions. When there was a mismatch between instructions and essay, i.e., when an essay with an instruction differed from a different writing task, ChatGPT-3.5 responded with the identification of the mismatch between instruction and essay, as shown in Figure 1.

Figure 1. *Response to mismatch.*



4.2. ChatGPT as Proofreader

During the interactions with ChatGPT-3.5 with requests to revise the IELTS essays, we found that ChatGPT-3.5 tended to re-write the essays rather than simply correcting the problematic areas. That is, ChatGPT changed the structure of sentences, the structure of paragraphs, and even the content of the essays.

Among all the modifications performed by ChatGPT-3.5 in the 10 selected essays, Lexical Resources was the most often revised descriptor (see Table 6).

Table 6. *Modifications to lexical resources.*

Strategy	Occurrences
change a word	179
add an adjective	8
add a phrase	7
correct spelling	3
add a clause	3
total	200

Among the recorded modifications of lexical resources, the most used strategy by ChatGPT-3.5 in revising lexical resources was to “change a word”, which was found 179 times in the revisions to the ten essays. Based on further analysis of these modifications, we found that the tool usually uses synonyms to replace original words. In most cases, the revised words were more infrequent or complex, as in changing the expression “some people dead” to “fatalities.” However, there were also occasions where the revision did not appear to significantly enhance the difficulty level of the words, as in changing “in my opinion” to “in my view.” Examples of word changes are shown in Table 7.

Table 7. *Word modifications by ChatGPT-3.5.*

Original	Revised
in our rather futuristic society	in today's rapidly evolving society
getting more interested	developing a keen interest
in my opinion	in my view
some people dead	fatalities
hometowns	homes and neighborhoods
help	assistance
have been drawn to the attention	has garnered the attention
thus	consequently

The second most revised descriptor was Cohesion and Coherence, with 50 occurrences identified in the revised 10 essays, as displayed in Table 8.

Table 8. *Modifications to Cohesion and Coherence.*

Strategy	Occurrences
add a topic sentence	11
add a connective	11
change a connective	9
restructure	8
add a conclusion	6
clarify reference	3
subject unification	1
add a recap	1
total	50

As shown in Table 8, the most used strategy for revising Cohesion and Coherence in the essays was "add a topic sentence" and "add a connective," with both appearing 11 times in the revisions of the 10 sample essays. "Add a topic sentence" refers to the original essay lacking an overall statement of key ideas at the beginning (or elsewhere) in a paragraph, in response to which ChatGPT-3.5 generated a topic sentence to make up for this deficiency. Examples of topic sentences added by ChatGPT-3.5 are displayed in Table 9.

Table 9. *Examples of topic sentences added by ChatGPT-3.5.*

1. "Raising a child is a profound responsibility that demands love, care, and readiness."
2. "Today, the scenario has undergone a profound transformation."
3. "This essay delves into the reasons behind this growing interest and explores various means by which individuals can research the history of their dwellings."

Regarding the Add Connective strategy, which comes under the Cohesion and Coherence descriptor, Table 10 shows the specific connective words that were added to the 10 selected essays.

Table 10. *Record of added connectives.*

Connective	Occurrences
furthermore	3
however	2
not only; but also	1
additionally	1
moreover	1
in turn	1
secondly	1
conversely	1
total	11

The descriptors of "Task Response" and "Grammar" recorded the same amounts of revisions, with 37 occurrences in total. Three strategies were identified by ChatGPT-3.5 under Task Response, namely "add details", "clarification", and "rationalization," as shown in Table 11. "Add details" refers to ChatGPT-3.5 adding new content to enrich the original text, and the added content is primarily not involved in the original essays. "Clarification" refers to revisions made by ChatGPT to present the original content more clearly. The difference between "Clarification" and "Add details" is that "Clarification" does not add new ideas but only chooses a clearer way to express the author's original idea. In the analysis of the 10 sample essays, "Add details" was found 19 times and "Clarification" 15 times.

The third strategy under "Task Response" is "Rationalization," which refers to providing a rationale for the writer's idea. In some cases, the writer uses strong but unsupported arguments that express ideas powerfully, as in "something must happen" or "it is never possible." In such cases, ChatGPT decreased the (unsupported) strength of the argument, thus enhancing the rationality of the idea, a strategy found on 3 occasions in the 10 essays.

Table 11. *Modifications under Task Response.*

Strategy	Occurrences
add detail	19
clarification	15
rationalization	3
total	37

Table 12 shows the strategies adopted by ChatGPT-3.5 to revise essays in terms of the Grammar descriptor.

Table 12. *Modifications to Grammar.*

Strategy	Occurrences
complication	14
change voice	10
change subject	3
word re-order	7
change sentence structure	4
total	38

The most used strategy was defined as "complication," which refers to grammar being made more complex. To distinguish "complication" from the other strategies under this descriptor, the criterion we chose was the enhancement of grammatical complexity. For example, in one essay, the original sentence "… my view is elaborated further" was revised to "I will elaborate on …." In this case, we classified the revision as "change voice" rather than "complication" since the level of grammatical complexity was not enhanced. An example of "complication" was found in another sentence from a sample essay, in which the original opening was "In this essay, I will try to discuss…" and the revised text was "…, which I will discuss in this essay." Here, the original simple sentence was combined with the previous sentence by transforming it into an attributive clause, which can be considered a step further in grammatical complexity.

Table 13 displays the distribution of the 14 occurrences of complications involving four types of revisions.

Table 13. *Complication.*

Strategy	Occurrences	Year of First Publication	
		Original	Revised
Change independent sentence to attributive clause	5	… and their levels of health and fitness are decreasing.	…, accompanied by a decline in overall health and fitness levels.
Change independent sentence to adverbial clause	5	…, as you do not have to go to a pharmacy …	…, sparing individuals the financial burden …
Change attributive phrase to parentheses	2	The smartphone connected with the internet opens up …	Smartphones, when connected to the internet, open up …
Change independent sentence to parentheses	2	Usually we have to pay around $30 for admissions.	The cost of entry, often around $30, can …

Table 14 displays the scores given by ChatGPT-3.5 to both original and revised essays, revealing a sharp difference between the two groups of scores.

Table 14. *Scores for revised essays given by ChatGPT-3.5.*

Descriptor	Original Essay	Revised Essay
Task Response	5.9	7.7
Coherence & Cohesion	5.7	7.8
Lexical Resources	5.5	7.8
Grammar	5.6	7.9
Overall Band	5.6	7.8

As Table 14 shows, although the grades given by ChatGPT-3.5 differ from the official scores, thus addressing our first research question, based on the scores given to the revised essays, it can be concluded that ChatGPT-3.5 was effective as a proofreader, at least to some extent. However, since the scores for the revised essays were given by ChatGPT-3.5 itself, the next step in the research should be to invite real IELTS examiners to evaluate the revised essays and compare their scores with the original essays.

An interesting phenomenon is that although the instruction to ChatGPT was to "revise the essay to a band 7 score," the tool generally revised all the essays to an average score of 7.8, which did not meet our requirement but exceeded the expected score.

5. DISCUSSION

The research found that an AES system such as ChatGPT-3.5 cannot be regarded as an ideal grader of IELTS exams since scores were generally lower than those given by official raters, with a significant gap in the grading outcomes. Thus, the inaccuracies in ChatGPT-3.5's grading outcomes might, at least for now, mitigate the concern the over total replacement of human raters or teachers (Warschauer *et al*., 2023). Moreover, the findings illustrate the difference in the scores generated by ChatGPT depends on whether or not an instruction was issued along with the essay inputs. The results imply that ChatGPT can read and consider instructions while assessing the essays. However, providing instructions does not make the scoring output more accurate but rather more different from the official scores. This indicates a limitation of ChatGPT-3.5 to take the writing instruction from the IELTS question we provided into appropriate consideration since our provision of this information did not help ChatGPT-3.5 grade more accurately. Moreover, the data showed no significant difference between having instructions input or not. Thus, ChatGPT can only be considered an inconsistent assessor, which makes it unsuited to what Yang and Dai (2015) call machine-assisted language testing. However, since the gap in average scores between ChatGPT and official scores was less than 0.5, ChatGPT can still be used as a supplementary tool in self-study, as in Huynh-Cam *et al*. (2024) and Mizumoto and Eguchi (2023), or a machine-assisted human rating.

As a proofreader, ChatGPT-3.5 showed comprehensive abilities in revising all the descriptors of the IELTS benchmark, as suggested in Stevenson and Phakiti (2014) about AEW tools. This finding is based on a qualitative perspective, with the researchers doing text analysis and manually coding the revisions. However, a much higher average score was given by ChatGPT itself after revising all the sample essays, a positive outcome that is in sharp contrast with the results from Bašić *et al*. (2023), who found GPT-3.0 to be ineffective in assisting students' essay writing. Three possible reasons for this finding can be suggested. The first may be the difference between GPT-3.0, the version used by Bašić *et al*. (2023), and ChatGPT-3.5, which was employed in this research. Second, even though GPT can revise essays, it may not be readily adopted by students, as He's (2016) study. Third, the essays in the study by Bašić *et al*. (2023) were not official exams and thus, unlike IELTS, had no acknowledged rubrics. Thus, the effectiveness of GPT-3.5 as a reliable proofreader can be attributed to the type of essays

under consideration as well as its use of popular rubrics such as that used by IELTS and similar exams. This finding aligns with ethical concerns raised by Chiu et al. (2023) regarding textual appropriation and plagiarism in academic writing. ChatGPT's improved performance with well-established, often-studied exams such as IELTS, which focus more on rhetorical strategies than the content itself, highlights potential risks as familiarity with these exams could make it easier for students to rely solely on AI to produce more accurate responses without truly engaging with the content or developing their writing skills.

Regarding ChatGPT-3.5's ability to revise English essays, there was a sharp difference with previous studies that denied the effectiveness of ChatGPT 2 or 3 (Bašić *et al*., 2023; Fyfe, 2022). This suggests two main reasons for the differences between the studies. One of the potential causes may be the gap between theoretical and practical research. Our study explored the effectiveness of ChatGPT from qualitative aspects through our interactions with the tool itself (see Fyfe, 2022; Kuhail *et al*., 2023; Pavlik, 2023) along with our analysis of the output. However, previous studies were mostly of a practical or empirical type, utilizing the tool with students and analyzing their performance (Huynh-Cam *et al*., 2024; Li, 2021; Mizumoto & Eguchi, 2023; Zhang, 2020). This methodological difference could thus be the cause of the inconsistency noted above. Another aspect, as noted above, could be the difference in the version of ChatGPT used, as previous studies investigated earlier versions. Thus we strongly recommend that future research adopt ChatGPT-3.5 (even ChatGPT-4 for the latest technology) in teachers' and students' practices.

The fact that we have experience of ChatGPT places us on an unusual path. For example, we were able to observe how global writing skills and language mechanics (Stevenson & Phakiti, 2014) and common L2 writing mistakes (Liang *et al*., 2023) could both be tackled by ChatGPT. For example, when ChatGPT pointed out issues regarding strong assumptions, we could identify how the way we express ideas might be misinterpreted, including ideas we often do not see as problematic but as enriching our texts. Moreover, we were able to verify what strategies students might have come across when choosing the suggestions given by ChatGPT (Barrot, 2023; Cotos, 2023; Huynh-Cam *et al*., 2024; Stevenson & Phakiti, 2014). Such strategies might inform pedagogical practices that aim to promote students' autonomy (Artiles Rodríguez *et al*., 2021; Barrot, 2023; Baskara, 2023; Chiu *et al*., 2023; Fyfe, 2022; Warschauer *et al*., 2023). They may also be useful in reducing teachers' essay correcting workload (Bai *et al*., 2022; Han *et al*., 2023; Li, 2021; Ranalli, 2018; Yang & Dai, 2015), particularly in the earlier phases of writing (such as drafting).

With regards to the limitations of this study, the coding of the proofreading, though monitored by a teacher, was conducted by two human researchers. Even though this has shown to provide high intercoder reliability, there may be some disputable points regarding categorizing the strategies used in revisions. Second, the sample involved only 23 essays, which may compromise the findings of our quantitative research. Furthermore, as we point out earlier in this paper, there should be another round of human raters, preferably IELTS raters, to assess the output of ChatGPT.

6. CONCLUSION

This study investigated two functions of ChatGPT-3.5 in addressing L2 writing. As a scoring system, ChatGPT-3.5 demonstrates the ability to provide referable scores but lacks the consistency needed to replace human raters entirely. Given the statistically significant gap between AI-generated scores and official rater scores, we should highlight the need for the cautious application of AI in grading high-stakes assessments. In contrast, as a proofreading tool, ChatGPT-3.5 shows significant potential, offering valuable revisions that help students improve lexical resources, cohesion, and overall writing quality. These findings suggest that while ChatGPT-3.5 may not yet be a solution for automated grading, it can effectively support teachers and students in the writing process, particularly in the formative stages. Our research

provides a reference to teachers and learners on how reliable and useful ChatGPT is, which in their future teaching and learning will act as a parameter for deciding whether to trust it or not or at least the extent of one's responsibility while using the tool. Future research should explore the integration of advanced AI versions of the tool in practical classroom applications in order to refine their reliability and maximize their pedagogical benefits. By addressing the limitations identified in this study, including the need for larger sample sizes and additional human rater evaluations, researchers can attempt to elucidate the role of AI tools in fostering autonomous learning environments.

Acknowledgments

We would like to express our gratitude to Yangfan Xu, who assisted us with the statistical analysis. This work is supported in part by the UIC Research Grant with No. of UICR0700033–22 at BNU–HKBU United International College, Zhuhai, China.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors.

Contribution of Authors

Xinming Chen: Study conception and design, data collection, analysis and interpretation of results, and manuscript preparation. **Ziqian Zhou:** Study conception and design, data collection, analysis and interpretation of results, and manuscript preparation. **Malila Prado:** Study conception and design, supervision of data collection, analysis and interpretation of results, and manuscript preparation.

Orcid

Xinming Chen https://orcid.org/0009-0008-6860-8633
Ziqian Zhou https://orcid.org/0009-0004-6937-1667
Malila Prado https://orcid.org/0000-0001-6281-6759

REFERENCES

Artiles Rodríguez, J., Guerra Santana, M., Aguiar Perera, V., & Rodríguez Pulido, J. (2021). Agente conversacional virtual: La inteligencia artificial para el aprendizaje autónomo. *Pixel-Bit, Revista de Medios y Educación*, *62*, 107–144. https://doi.org/10.12795/pixelbit.86171

Bai, J.Y.H., Zawacki-Richter, O., Bozkurt, A., Lee, K., Fanguy, M., Cefa Sari, B., & Marin, V.I. (2022, September). Automated Essay Scoring (AES) Systems: Opportunities and challenges for open and distance education. Tenth Pan-Commonwealth Forum on Open Learning. https://doi.org/10.56059/pcf10.8339

Barrot, J.S. (2023). Using ChatGPT for second language writing: Pitfalls and potentials. *Assessing Writing*, *57*, 100745. https://doi.org/10.1016/j.asw.2023.100745

Bašić, Z., Banovac, A., Kruzic, I., & Jerkovic, I. (2023). *Better by you, better than me: ChatGPT3 as writing assistance in student essays*. https://doi.org/10.48550/ARXIV.2302.04536

Baskara, R. (2023). Integrating ChatGPT into EFL writing instruction: Benefits and challenges. *International Journal of Education and Learning*, *5*(1), 44-55. https://doi.org/10.31763/ijele.v5i1.858

Boa Sorte, P., Farias, M.A. de F., Santos, A. E., Santos, J. do C.A., & Dias, J.S. dos S.R. (2021). Artificial intelligence in academic writing: What is the CPT-3 algorithm? *Revista EntreLinguas*, *7*, e021035.

Chiu, T.K.F., Xia, Q., Zhou, X., Chai, C.S., & Cheng, M. (2023). Systematic literature review of opportunities, challenges, and future research recommendations of artificial intelligence

in education. *Computers and Education: Artificial Intelligence*, *4*, 100118. https://doi.org/10.1016/j.caeai.2022.100118

Cotos, E. (2023). Automated feedback on writing. In O. Kruse, C. Rapp, C.M. Anson, K. Benetos, E. Cotos, A. Devitt, & A. Shibani (Eds.), *Digital writing technologies in higher education* (pp. 347–364). Springer International.

Crawley, M.J. (2012). *The R book*. John Wiley & Sons.

Dergaa, I., Chamari, K., Zmijewski, P., & Ben Saad, H. (2023). From human writing to artificial intelligence generated text: Examining the prospects and potential threats of ChatGPT in academic writing. *Biology of Sport*, *40*(2), 615-622. https://doi.org/10.5114/biolsport.2023.125623

Fryer, L.K., Coniam, D., Carpenter, R., & Lăpușneanu, D. (2020). Bots for language learning now: Current and future directions. *Language Learning & Technology*, *24*(2), 8–22. http://hdl.handle.net/10125/44719

Fyfe, P. (2022). How to cheat on your final paper: Assigning AI for student writing. *AI & Society*, *38*, 1395–1405. https://doi.org/10.1007/s00146-022-01397-z

Han, J., Yoo, H., Kim, Y., Myung, J., Kim, M., Lim, H., Kim, J., Lee, T. Y., Hong, H., Ahn, S.-Y., & Oh, A. (2023). RECIPE: How to integrate ChatGPT into EFL writing education. *Proceedings of the Tenth ACM Conference on Learning @ Scale*, 416–420. https://doi.org/10.1145/3573051.3596200

He, H. (2016). A survey of EFL college learners' perceptions of an on-line writing program. *International Journal of Emerging Technologies in Learning (Online)*, *11*(4), 11-15. https://doi.org/10.3991/ijet.v11i04.5459

Huang, H.-W., Li, Z., & Taylor, L. (2020). The effectiveness of using Grammarly to improve students' writing skills. *Proceedings of the 5th International Conference on Distance Education and Learning*, 122–127. https://doi.org/10.1145/3402569.3402594

Huang, W., Hew, K.F., & Fryer, L.K. (2022). Chatbots for language learning—Are they really useful? A systematic review of chatbot-supported language learning. *Journal of Computer Assisted Learning*, *38*(1), 237-257. https://doi.org/10.1111/jcal.12610

Huynh-Cam, T.-T., Agrawal, S., Bui, T.-T., Nalluri, V., & Chen, L.-S. (2023). Enhancing the English writing skills of in-service students using Marking Mate automated feedback. *Asia Pacific Education Review*, *25*(2), 459–474. https://doi.org/10.1007/s12564-023-09904-7

Kohnke, L., Moorhouse, B.L., & Zou, D. (2023). ChatGPT for language teaching and learning. *RELC Journal*, *54*(2), 537–550. https://doi.org/10.1177/00336882231162868

Kuhail, M.A., Alturki, N., Alramlawi, S., & Alhejori, K. (2023). Interacting with educational chatbots: A systematic review. *Education and Information Technologies*, *28*(1), 973–1018. https://doi.org/10.1007/s10639-022-11177-3

Lee, S.M. (2023). The effectiveness of machine translation in foreign language education: A systematic review and meta-analysis. *Computer Assisted Language Learning*, *36*(1-2), 103–125. https://doi.org/10.1080/09588221.2021.1901745

Li, Z. (2021). Teachers in automated writing evaluation (AWE) system-supported ESL writing classes: Perception, implementation, and influence. *System*, *99*, 102505. https://doi.org/10.1016/j.system.2021.102505

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns*, *4*(7), 100779. https://doi.org/10.1016/j.patter.2023.100779

Miles, M.B., & Huberman, A.M. (1994). *Qualitative data analysis: An expanded sourcebook*. Sage.

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), 100050. https://doi.org/10.1016/j.rmal.2023.100050

Open AI. (2022). *Introducing ChatGPT*. https://openai.com/blog/chatgpt

Pavlik, J.V. (2023). Collaborating with ChatGPT: Considering the implications of generative artificial intelligence for journalism and media education. *Journalism & Mass Communication Educator*, *78*(1), 84–93. https://doi.org/10.1177/10776958221149577

Prado, M.C.A., & Huggins, T.J. (2023). Technological approaches to student participation while studying the history of psychology in an EMI institution. In J. Corbett, E.M.Y. Yan, J. Yeoh, & J. Lee (Eds.), *Multilingual Education Yearbook 2023* (pp. 49–69). Springer International. https://doi.org/10.1007/978-3-031-32811-4_4

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

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

Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. *Assessing Writing*, *19*, 51–65. https://doi.org/10.1016/j.asw.2013.11.007

Test Statistics. (2022). IELTS. https://ielts.org/researchers/our-research/test-statistics#Demographic

Uysal, H.H. (2010). A critical review of the IELTS writing test. *ELT Journal*, *64*(3), 314–320. https://doi.org/10.1093/elt/ccp026

Warschauer, M., Tseng, W., Yim, S., Webster, T., Jacob, S., Du, Q., & Tate, T. (2023). The affordances and contradictions of AI-generated text for writers of English as a second or foreign language. *Journal of Second Language Writing*, *62*, 101071. https://doi.org/10.1016/j.jslw.2023.101071

Yang, X., & Dai, Y. (2015). An empirical study of college English autonomous writing teaching mode based on www.pigai.org. *Technology Enhanced Foreign Language Education, 162*(02), 17-23. (Translated from Chinese) https://doi.org/10.3969/j.issn.1001-5795.2015.02.003

Zhang, Z.V. (2020). Engaging with automated writing evaluation (AWE) feedback on L2 writing: Student perceptions and revisions. *Assessing Writing*, *43*, 100439. https://doi.org/10.1016/j.asw.2019.100439

Zhou, Z., & Prado, M. (2024). A corpus-based comparative study of readability of passages in compulsory Chinese English textbooks and exams for middle school students. *Proceedings of the 13th Int. Conf. on Educational and Information Technology.* pp. 279-83. http://doi.org/10.1109/ICEIT61397.2024.10540975