The Reliability of using ChatGPT in Rating EFL Writings

OPEN ACCESS

Manuscript ID:
EDU-2024-12047855

Volume: 12

Issue: 4

Month: September

Year: 2024

P-ISSN: 2320-2653

E-ISSN: 2582-1334

Received: 20.06.2024

Accepted: 20.08.2024

Published: 01.09.2024

Citation:
Yang, Y. (2024). The Reliability of using ChatGPT in Rating EFL Writings. *Shanlax International Journal of Education, 12*(4), 49-59.

DOI:
https://doi.org/10.34293/education.v12i4.7855

Yang Yang
Southwest University, China
https://orcid.org/0000-0003-3114-0682

Abstract
This paper explores the reliability of using ChatGPT in evaluating EFL writing by assessing its intra- and inter-rater reliability. Eighty-two compositions were randomly sampled from the Written English Corpus of Chinese Learners. These compositions were rated by three experienced raters with regard to 'language', 'content', and 'organization'. The writing samples were also rated by ChatGPT twice over some time, and the average scores were calculated. Independent samples t-test was conducted to compare the average scores given by ChatGPT and human raters. Pearson correlation analyses were conducted between the two sets of overall scores given by ChatGPT to calculate the intra-rater reliability, as well as between average scores given by ChatGPT and human raters for inter-rater reliability. The results of comparative analysis shows that ChatGPT may be used for evaluating EFL essays, as the scores are similar to those provided by reliable human raters. However, the result of correlation analyses shows that the intra-rater reliability of ChatGPT is not high enough to be acceptable, r=0.575, p<0.01 and the strength of the inter-rater reliability is moderate as well, r=0.508, p<0.01. Besides, there is no significant relationship between their average scores on 'organization' of the writings, r=0.181, p>0.05. Thus, it can be concluded that ChatGPT is not a reliable tool to rate and score EFL writings using the prompt in this study. One of the possible reasons for the unreliability of ChatGPT as a rater of EFL writing seems to be related to scoring for the 'organization' of the essay. These findings imply that while ChatGPT has potential as an evaluative tool, its current limitations, particularly in assessing organization, must be addressed before it can be reliably used in educational settings.

Keywords: Reliability, ChatGPT, Rating, Writing Evaluation, EFL Writing

Introduction

ChatGPT (Chat Generative Pre-trained Transformer) is a natural language processing tool driven by artificial intelligence (AI) technology launched by OpenAI, an American AI research laboratory. By using transformer neural network architecture and connecting a large number of corpora to train the model, it has the ability of language understanding and text generation. As a chatbot, it is designed to mimic human conversation and engage with users (King, 2023). Unlike previous chatbots, such as Siri by Apple or Meena by Google, ChatGPT is a generative model, which means it can generate new data, as opposed to only classifying or predicting based on input data (Pavlik, 2023).

Within a short time since ChatGPT 3.0's release on November 30, 2022, there has been a flurry of research on its use in education (Baidoo-Anu & Owusu Ansah, 2023; Halaweh, 2023; Hong, 2023; Pavlik, 2023; Rudolph et al., 2023; Zhai, 2022; Zimmerman, 2023). It is claimed that personalized and interactive learning, formative assessment techniques, and other advantages of ChatGPT in education are just a few (Baidoo-Anu & Owusu Ansah, 2023). It is also reported that this ChatGPT can aid in medical education and help with clinical decision-making, as it is capable of providing precise responses in medical licensure exams (Kung et al., 2023).

In the field of EFL writing education, including writing instruction, learning, practicing, and assessment, it has been reported that ChatGPT can help to improve the writing quality of EFL learners by correcting grammatical and stylistic errors and making the writing more comprehensible (Atlas, 2023; Kohnke et al., 2023). It can aid in writing research papers, as it can introduce writers to new research topics and provide them with resources and information on a particular topic (Kasneci et al., 2023). It can also assist writing teachers with writing instruction (Rudolph et al., 2023). In addition, it is obvious that ChatGPT will be considered to rate EFL learner's written products since it is possible to insert a learner's writing in the chat box and ask ChatGPT to rate it and give score and comments based on the requirement specified by the teacher or instructor. Though it is claimed that AI-powered chatbots can conduct formative language assessment and provide immediate feedback (Huang et al., 2022; Kuhail et al., 2023), there is little research on the application of ChatGPT to EFL writing assessment.

There are already many software or systems for Automated Writing Evaluation (AWE) or Automated Essay Scoring (AES) that are used, such as *IntelliMetric* (Rudner et al., 2006), *E-rater* (Burstein et al., 2004), and *Intelligent Essay Assessor* (Landauer, 2003). *IntelliMetric* is constructed upon an amalgamation of AI, natural language processing, and statistical methodologies. Empirical evidence suggests that the agreement between scores designated by *IntelliMetric* and those allocated by human evaluators is notably high (Sathyabalan & Christian, 2022). The *E-rater* system employs a contemporary statistical and rule-based approach facilitating the examination of syntax, morphology, and semantics (Burstein et al., 2013). Owing to its documented reliability and validity (Attali & Burstein, 2004), the E-rater has been formally integrated alongside human evaluators in high-stakes tests, including the Graduate Record Examination (GRE) and the Test of English as a Foreign Language (TOEFL) (Mizumoto & Eguchi, 2023).

In China, there are also intelligent assessment systems or platforms that are specially designed for Chinese EFL learners, such as *Pigai*[1], *iTEST*[2], or *iWrite*[3], among which, the *Pigai* system is the most popular and influential one. 'Pigai' is the Pinyin of 批改 (correction). The Pigai system, a product of the National Language Intelligence Center of China, emerged in 2011 as a commercialized online evaluation platform, specifically tailored for Chinese EFL learners (Bai & Hu, 2017). Empirical evidence suggests that Pigai holds the potential to enhance students' writing proficiency through the provision of insightful and beneficial feedback (Wu, 2018). Nevertheless, its limitations manifest in its inability to discern content-associated attributes within an essay (Li, 2014).

Though these AWE software or systems have advantages of 'time and cost saving' and 'efficiency in grading and providing feedback' (Zhang, 2021), compared to interactive AI language models like ChatGPT, a significant drawback of theirs is the inability to interact with users. That is, with the scores or feedback given by these AWE systems, language learners or instructors cannot engage them with specific follow-up inquiries. Armed with this inherent advantage, it is suggested that ChatGPT could be utilized for AWE or AES applications (Essel, 2023).

However, as an emerging technology, how reliable ChatGPT is in evaluating EFL writing remains to be seen since 'scoring consistency is an important aspect of evaluating the AES system' (Mizumoto & Eguchi, 2023). In the existing relevant literature, there are hardly any studies specifically examining the reliability of applying ChatGPT to the rating of EFL writing, with a few exceptions: Mizumoto and Eguchi (2023) explored the intra-rater reliability of ChatGPT grading but failed to introduce a reliable human scoring reference to further investigate its inter-rater reliability. To fill this gap, this study aims to investigate the reliability, including the intra- and inter-rater reliability, of ChatGPT's rating on EFL writing by taking reliable manual ratings as a reference. The research questions are as follows:

- What is the difference between average scores given by ChatGPT and human raters?
- What is the intra-rater reliability of ChatGPT's rating on EFL writing?
- What is the inter-rater reliability between ChatGPT's rating and reliable manual rating on EFL writing?

[1] http://www.pigai.org
[2] https://itestcloud.unipus.cn
[3] http://iwrite.unipus.cn/

This study addresses a critical gap in the literature by systematically examining the reliability of ChatGPT as a tool for evaluating EFL writing. Given the widespread use of AI-powered tools in educational settings, understanding ChatGPT's reliability in this context has significant implications for the future of automated language assessment. By providing empirical evidence on both intra- and inter-rater reliability, this research offers valuable insights into the potential and limitations of ChatGPT as an evaluative tool. The findings could inform educators, policymakers, and developers about the viability of integrating ChatGPT into EFL writing assessment, ultimately contributing to more effective and interactive assessment methods. This research may also stimulate further studies and innovations in the application of AI in educational assessment, enhancing the overall quality and fairness of language testing for learners worldwide.

Methods

Data Collection

The writing samples used in this study were extracted from the Written English Corpus of Chinese Learners (WECCL) (Wen et al., 2008). WECCL comprises 4,950 timed and un-timed compositions written by English majors and a fraction of non-English majors from more than 20 universities all over the country. These compositions can well reflect the writing performance of Chinese university EFL learners (Wen et al., 2008), and much research (e.g., Tang & Cao, 2021; Yan, 2019; Yan & Li, 2019; Tao & Yan, 2020) has been conducted based on this corpus. In this corpus, there are 270 expository compositions written by 270 Chinese EFL learners within a time limit of 30 minutes. Eighty-two compositions were randomly sampled by using the *Random Integer Set Generator*[4] (Yang & Zheng, 2024).

The sample size was calculated by using G*Power (Faul et al., 2009; Faul et al., 2007) to conduct a priori analysis. By choosing tail(s) as two, inputting parameters of a conventionally medium effect size of 0.3 (Cohen, 1988), significance level of 0.05, and a conventionally high enough power of 0.8, the result of the priori analysis showed that at least 82 samples were needed in a correlation analysis to reach the above-mentioned effect size and power.

Data Processing

The 82 compositions were rated by three raters on aspects of language (40%), content (30%), and organization (30%), and the total score was the sum of the three parts. These three raters are university instructors with extensive experience in teaching English Writing courses to university students and in evaluating university-level English writing, including grading the College English Test band 4 (CET-4) and CET-6 writing sections and English writing course assignments. Prior to rating these English writing samples, the raters underwent systematic training. They assessed the writing samples according to the criteria used for the CET-6 English writing exam, with the difference that the original scoring scale was converted to a percentage scale. Then, the average scores of the total score and average scores of each aspect from the three raters were calculated. Part of the scores given by the three raters is shown in Table 1.

[4] https://www.random.org/integer-sets

Table 1 Rating Scores from the Three Manual Raters

ID	Rater A	Rater B	Rater C	Language*	Content*	Organization*	Average score
WEXP0001	77.60	81.67	74.00	32.76	22.67	22.33	77.76
WEXP0002	57.20	66.67	58.00	24.76	20.53	15.33	60.62
WEXP0003	74.27	80.33	75.33	27.51	22.13	27.00	76.64
WEXP0004	83.47	83.33	80.00	34.00	24.40	23.87	82.27
WEXP0005	75.27	73.93	63.00	31.60	23.27	15.87	70.73
…							
WEXP0266	64.60	75.60	74.13	26.58	22.67	22.20	71.44

Note: *The score here represents the average score given by three raters in this aspect

The inter-rater reliability analysis between scores from every two raters was calculated. The results showed that they have significant (p<0.01) and high (Bachman, 2004; Carr, 2011; Guilford, 1973) inter-rater reliabilities since the corresponding correlation coefficients were from 0.710 to 0.785 (Yang et al., 2023b). The results of the reliability analysis is shown in Table 2.

Table 2 Inter-Rater Reliability Between Every Two Raters

		Rater A	Rater B	Rater C
Rater A	Pearson Correlation	1		
Rater B	Pearson Correlation	.720**	1	
Rater C	Pearson Correlation	.710**	.785**	1

Note: **Correlation is significant at the 0.01 level (2-tailed)

Then, the 82 EFL compositions were input in the chat box of ChatGPT one by one for rating. After each input, the rating with scores and comments was given by ChatGPT in seconds. The next day, the same 82 writings were rated by ChatGPT again with the same prompts to obtain another set of scores. Then, the average of ChatGPT's two ratings is calculated. The prompt is as follows:

#WEXP0XXX

'…' (EFL writing of the above ID)

The above is a piece of writing by a Chinese university student, who is allowed to write a report of 150-180 words in 30 minutes about the development of KFC and MacDonald's over a ten-year period in China with the reference of the following table. Please rate the writing from aspects of 'language' (40 marks), 'content' (30 marks), and 'organization' (30 marks) and give marks for each aspect and the overall mark.

Table. Number of stores of KFC and MacDonald's over a ten-year period in China

Year	*1994*	*1995*	*1996*	*1997*	*1998*	*1999*
	2000	*2001*	*2002*	*2003*	*2004*	
KFC	*45*	*72*	*131*	*216*	*292*	*327*
	400	*534*	*902*	*1000*	*1200*	
MacDonald's	*6*	*11*	*53*	*122*	*145*	*195*
	214	*353*	*543*	*573*	*600*	

Although ChatGPT also offers a paid version in the form of GPT-4 and GPT-4o, this study utilizes the free ChatGPT-3.5 version. The choice of version 3.5 is motivated by its broader user base, making the results of this study more pertinent to a larger audience. As for whether employing version 4 and 4o would yield different outcomes, subsequent investigations will be conducted in future research.

Data Analysis

To answer the first research question, an independent-sample t-test was conducted to find out if there was any difference between average scores given by ChatGPT and human raters. Also, the standard deviation, minimum, and maximum values in the average scores were compared and analyzed.

To answer the next two research questions, two correlation analyses were conducted. First, a correlation between the two sets of overall scores of ChatGPT's rating on EFL writings was analyzed to find out the consistency of ChatGPT's ratings over some time with the same writings. Then, regarding ChatGPT as a 'rater', a correlation between the average scores given by ChatGPT and human raters was analyzed to investigate their inter-rater reliability. If both correlation coefficients were larger than 0.7, the reliability would be conventionally acceptable (Bachman, 2004; Carr, 2011; Guilford, 1973), and it would be safe to conclude that it is reliable for ChatGPT in rating EFL writings. Then, a series of follow-up correlation analyses between ChatGPT's and manual ratings on each aspect, namely 'language', 'content', and 'organization', were conducted to find out which aspect of the EFL writing ChatGPT rates more reliably.

When conducting the correlation analyses, the Pearson Correlation Coefficients were chosen over the Spearman Correlation Coefficients due to the large sample size of the study and the numerical nature of both variables. According to the skewness and kurtosis values shown in Table 3, they were between ±2, so the data were regarded as normal (George & Mallery, 2003). Besides, the compositions were randomly sampled. Thus, the assumptions of independent-sample t-test and correlation analysis were met, including random sampling, normality of data distribution, adequacy of sample size, numeric measurement, et cetera.

All the data used in this study, including the randomly sampled compositions, ChatGPT's comments, and scores given by ChatGPT and human raters, have been reviewed and published on *Mendeley Data* (Yang et al., 2023a), and the present study can be replicated with these data.

Result and Discussion

Research Question 1

This section answers the first research question: are there significant differences between the scores given by ChatGPT and human raters? Table 3 shows descriptive statistics of manual and ChatGPT's rating on the 82 EFL writings, and Table 4 shows the result of the independent-samples t-test comparing the average scores given by human raters and ChatGPT.

As Table 3 shows, the average scores of manual and ChatGPT's ratings are similar (71.37 and 73.45), and the result of the independent-samples t-test in Table 4 shows that there is no significant difference between them, t(148.595) = -1.338, p=0.183>0.05. A primary conclusion can be drawn that ChatGPT may be used for evaluating EFL essays, as the scores are similar to those provided by reliable human raters.

Table 3 Descriptive Statistics of Manual and ChatGPT's Rating on EFL Writings*

	N	Min	Max	Mean	SD	Skewness	Kurtosis
Mannual_Average_score	82	50.49	89.76	71.37	8.32	-0.169	-0.391
ChatGPT1_Overall_score	82	40.00	94.00	73.23	12.98	-0.615	-0.125
ChatGPT2_Overall_score	82	37.00	93.00	73.66	12.58	-0.806	0.206
ChatGPT_Average_score	82	41.00	91.50	73.45	11.34	-0.674	0.205

Note: *Scores of 100 are the maximum, with 40 for language, 30 for content, and 30 for organization

Table 4 Result of the Independent-Samples t-test

	Equal variances	Levene's Test		t-test		
		F	Sig.	t	df	Sig. (2-tailed)
Score	Assumed	5.752	.018	-1.338	162	.183
	Not assumed			-1.338	148.595	.183

However, the standard deviation of ChatGPT's average scores (SD=11.34) is much larger than that of manual rating (SD=8.32). Besides, the minimum average score given by ChatGPT (41) is much smaller than that of the manual rating (50.49), and the maximum of that given by ChatGPT (91.5) is larger than that of the manual rating (89.76). With regard to the overall score distribution, ChatGPT's assigned scores are more dispersed compared to those given by human raters. This indicates that ChatGPT is more inclined to award higher scores to well-composed essays and, conversely, is willing to assign lower scores to less competent compositions. If ChatGPT were to be likened to a human evaluator, it could be said that ChatGPT is 'bolder' than human raters when rating EFL writings. For instance, ChatGPT gave Writing #WEXP0220 only 37 marks out of 100 in its second rating and the comment 'It's difficult to follow your ideas', while it is given an average of 62.64 by human raters. On the contrary, human raters tend to be more conservative and save 'faces' for test-takers. ChatGPT's scores are more dispersed than those of human raters primarily because the AI operates purely based on algorithmic criteria without the emotional or contextual considerations that human raters might apply. Human raters often show a tendency to cluster their scores within a narrower range to avoid extreme judgments, possibly due to concerns about fairness or empathy towards the student. In contrast, ChatGPT evaluates based solely on the data it was trained on, leading to a more literal interpretation of the quality of the writing, which results in a wider distribution of scores. This difference aptly highlights the fundamental distinction between human raters and machine scoring: machines are devoid of emotions, whereas human raters might consider the personal feelings of the test-taker when faced with a poorly written essay. They would assign a low score based on rubrics, but not so low as to embarrass the student.

While ChatGPT's objective approach can highlight differences in writing quality more distinctly, it may also lead to scores that seem too harsh or lenient compared to human evaluations. This suggests that while AI tools like ChatGPT can support educational assessments, they should be used in conjunction with human oversight to ensure the scores are both fair and contextually appropriate.

Research Question 2

The second research question, whether ChatGPT has intra-rater reliability for scoring EFL compositions, was addressed by examining the correlation between two sets of scores provided by ChatGPT.

Table 5 Correlation Coefficients of the Intra-Rater Reliability

		Pearson Correlation / Sig. (2-tailed)			
		ChatGPT's 2ed rating			
		Overall score	Language	Content	Organization
ChatGPT's 1st rating	Overall score	0.575**/0.000			
	Language		0.545**/0.000		
	Content			0.388**/0.000	
	Organization				0.497**/0.000

Note: **Correlation is significant at the 0.01 level (2-tailed).

Table 5 shows the correlation coefficients of the intra-rater reliability, which indicates that there is a significant and positive relationship between the two sets of overall scores given by ChatGPT, r(162)=0.575, p<0.01. However, according to Carr's (2011) rule of thumb, the strength of the correlation is moderate. It indicates that it is not highly consistent for ChatGPT to rate the same EFL writing samples twice over some time. In addition, the correlations of specific scores on the three aspects of EFL writing are also either low or moderate. It can be concluded that there is no acceptable intra-rater reliability for ChatGPT as a rater of EFL writing.

Research Question 3

The third research question, which asks if ChatGPT possesses inter-rater reliability for rating EFL writings, was answered by analyzing the correlation between scores from ChatGPT and those from a reliable human rater.

Table 6 Correlation Coefficients of the Inter-Rater Reliability

		Pearson Correlation / Sig. (2-tailed)			
		Manual rating			
		Average score	Language	Content	Organization
ChatGPT's rating	Average score	0.508**/0.000			
	Language		0.364**/0.001		
	Content			0.487**/0.000	
	Organization				0.181/0.103

Note: **Correlation is significant at the 0.01 level (2-tailed)

For the inter-rater reliability, the results of correlation analysis in Table 6 also show a significant and positive association between average scores given by ChatGPT and human raters, r(162)=0.508, p<0.01. As suggested by Carr (2011), the strength of the correlation is not acceptably high enough to claim the reliability of ChatGPT in rating EFL writings by taking reliable manual ratings as a reference. For the 'language' and 'content' aspects, the correlation coefficients showed that there is a low or moderate relationship between ChatGPT's and manual ratings. Understandably, language part, such as lexical and syntactic aspects, has always been highly correlated with the writing quality (Yang et al., 2022a, 2022b). However, 'organization' aspect, the result indicated that there is no significant association between them (p=0.103>0.05). It can be concluded that ChatGPT's rating on EFL writing is not consistent with the reliable manual rating, and it performs poorly when rating the 'organization' of the writing.

Here are some possible reasons for the unreliability of ChatGPT in EFL writing evaluation. The first possible reason is the weakness of ChatGPT in rating the 'organization' of EFL writing. Understandably, it is more difficult for software or even AI tools to evaluate 'organization' than it is to rate 'language' or 'content' of a composition. It is also reported that online grammar checkers, such as *Grammarly*[5], *ProWritingAid*[6], *Ginger*[7], and *Gram-marCheck*[8], can help correct language mistakes, 'but still they may not yet be optimum' in improving the organization of writing (Perdana & Farida, 2019). Thus, when using ChatGPT to rate EFL writings, it can be considered to add an explicit rubric for 'organization' to the prompt. For example, 'if the composition has 'wide range of explicit text organizational devices on essay and paragraph levels' (Ghalib & Al-Hattami, 2015), the aspect of 'organization' can be scored high'.

Another reason might be the different understanding of the cultural or contextual nuances of the writing task. Although ChatGPT was given as much information as possible, such as word count requirement, time limit, topic, and tabular data in the instructions of the writing task, it may not understand the context of the task as fully as a human rater (Taecharungroj, 2023), which may influence its rating. Finally, it might be caused by the limited training data. Though it seems that ChatGPT can answer any questions, it was not designed specifically to rate EFL writing like other AWE software or systems. If more training data about writing samples and reliable manual ratings are given, it may perform better.

However, the above conclusions of this study are drawn only based on quantitatively analyzing scores given by ChatGPT and human raters, but without qualitative analysis of the comments given by ChatGPT. If carefully examining the comments, it can be seen that most of them are reasonable. In terms of 'language', it can analyze the grammatical accuracy, vocabulary range, and use of idiomatic expressions in the writing. It also can assess the accuracy and relevance of the 'content' in the writing. Finally, with regard to 'organization', ChatGPT can assess the structure, coherence, and logical progression of the ideas in the writing. To reiterate, one of the advantages of using ChatGPT, compared to other AWE software is that it can interact with users, such as language teachers, instructors, or learners, for further inquiry on its comments when the users are confused about the comments given by ChatGPT. This is important since it is unknown to what degree language learners can understand automated feedback generated by AWE software (Zhang, 2021).

The conclusions of this study conflict with some existing related research findings. For instance, Mizumoto and Eguchi (2023) reported that AES using GPT can achieve a certain level of accuracy and acceptable intra-rater reliability. The discrepancy in the results between the two studies may arise from the use of different types of variables of writing scores. This study employed a scoring system based on a total of 100 points, averaging the scores of three human raters, and rendering the data as continuous variables. In contrast, Mizumoto and Eguchi (2023) utilized the IELTS band scoring criteria, with scores ranging from 0 to 9 across ten levels, and further categorized the scores into three broad tiers, low, medium, and high. Such data belongs to ordinal data, and this coarse granularity of data processing may reduce the informational content of the data. In other words, the precise score system based on a total of 100 points employed in this study might elevate the standards and complexity of both intra- and inter-rater reliability, leading to inconsistencies with previous research conclusions.

Although Mizumoto and Eguchi (2023) claim that AES using GPT possesses a certain level of reliability, they also acknowledge that 'it still falls short of achieving perfect agreement with human raters', and 'therefore, it should be used in conjunction with human evaluation' (p. 10). As a result, regarding the application of AI language models like ChatGPT and AWE/AES systems to score EFL writing, this study holds a view similar to that of previous scholars. That is to say, these systems can merely function as supplementary instruments and are not equipped to supplant human evaluators or in-class educators (Attali et al., 2013; Mizumoto & Eguchi, 2023; Warschauer & Ware, 2006).

[5] https://grammarly.com

[6] https://prowritingaid.com

[7] https://www.gingersoftware.com

[8] https://www.grammarcheck.net

Conclusion

With ChatGPT's release and popularity, some researchers, especially ones in the field of language assessment, are wondering whether it can be used in EFL writing evaluation. Trying to find out the reliability of ChatGPT in rating EFL writing, this study investigated its intra-rater reliability as well as the inter-rater reliability between ChatGPT's and reliable manual rating.

Based on scores and comments given by ChatGPT, it seems 'bolder' when rating the compositions, while human raters tend to be conservative and save 'faces' for test takers since the scores given by ChatGPT for poor compositions are smaller than that by human raters, and the distribution of the scores given by ChatGPT span a bigger range. Statistically, there is no significant difference in average scores given by ChatGPT and reliable human raters, suggesting that it may be used for evaluating EFL compositions. However, the result of the correlation analysis of intra-rater reliability indicates that it is not consistent for ChatGPT to rate the same writings over some time. The result of the correlation analysis of inter-rater reliability shows that scores of ChatGPT's rating is not highly correlated with that of reliable human raters. It can be concluded that it is not reliable to use ChatGPT to score the EFL writings. Even so, the great potential of ChatGPT in foreign language education, including foreign language teaching, learning, and assessment, cannot be completely dismissed.

One implication of the findings of this study is that EFL writing assessors or raters need to be cautious about scores of EFL writing given by ChatGPT. At the same time, they should pay attention to whether students' essays were written with the help of ChatGPT because the high-quality essays generated by ChatGPT can pass the detection of plagiarism checking software (Khalil & Er, 2023; Susnjak, 2022). Fortunately, many applications are being developed that can detect AI-generated text, such as *GPTZero*[9]. On the other hand, language instructors cannot completely ignore ChatGPT like AI-powered chatbots because it has been integrated into today's language education. In reaction to the rapid growth in digital technology, they should welcome them with open arms rather than avoid them. Due to the necessity of online instruction during the COVID-19 pandemic, language teachers have improved their digital literacy (Moorhouse, 2023), yet, it is claimed they still need the skills necessary to use ChatGPT effectively (Kohnke et al., 2023).

To advance the application of ChatGPT in educational assessment, future research should explore several critical areas. First, investigating how ChatGPT's feedback can be optimized to address different learning styles and needs could significantly enhance its utility in personalized education. This includes examining its effectiveness in providing tailored feedback for diverse student populations and subject areas. Additionally, studying the integration of ChatGPT with other educational technologies, such as adaptive learning platforms and learning management systems, could yield insights into creating more cohesive and supportive learning environments. Practical applications should also focus on developing robust frameworks for utilizing ChatGPT in real-time classroom settings, ensuring its feedback is actionable and aligned with pedagogical goals. Research into the ethical implications and biases inherent in ChatGPT's assessments is crucial for ensuring fairness and transparency. Lastly, longitudinal studies on the impact of ChatGPT-driven assessments on student performance and motivation would provide valuable data on its effectiveness and areas for refinement. These efforts collectively will help harness ChatGPT's potential to enhance educational outcomes and provide equitable support for diverse learners.

This study includes some limitations. It only quantitatively investigated scores of ChatGPT's rating, but without qualitatively examining comments and feedback given by it, which can be considered in future research. In addition, it remains uncertain whether using ChatGPT 4 and 4o to rate EFL writing would yield different results. Finally, though enough information is given to ChatGPT when it is asked to rate a composition, it has not been trained before. The working mechanism of large-scale AI language models like ChatGPT involves accomplishing new tasks after extensive pre-training; the more training data, the better the performance. Therefore, subsequent research could consider initially 'feeding'

[9] https://gptzero.me

ChatGPT with EFL writing and their corresponding human evaluations for its learning, and then test its reliability in scoring unseen writing.

References

Atlas, S. (2023). *ChatGPT for Higher Education and Professional Development: A Guide to Conversational AI.*

Attali, Y., & Burstein, J. (2004). Automated essay scoring with e-rater® v. 2.0. *ETS Research Report Series*, *2004*(2), i-21.

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

Bachman, L. F. (2004). *Statistical Analyses for Language Assessment*. Cambridge University Press.

Bai, L., & Hu, G. (2017). In the face of fallible AWE feedback: How do students respond?. *Educational Psychology*, *37*(1), 67-81.

Baidoo-Anu, D., & Owusu Ansah, L. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI*, *7*(1), 52-62.

Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The criterion online writing service. *AI Magazine*, *25*(3), 27-36.

Burstein, J., Tetreault, J., & Madnani, N. (2013). The e-rater® automated essay scoring system. In M. D. Shermis & J. Burstein (Eds.), *Handbook of Automated Essay Evaluation: Current Applications and New Directions* (pp. 55-67). Routledge.

Carr, N. T. (2011). *Designing and Analyzing Language Tests*. Oxford University Press.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Erlbaum.

Essel, H. (2023). 7 things you should know about ChatGPT. *BELI*.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149-1160.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175-191.

George, D., & Mallery, P. (2003). *SPSS for Windows Step by Step: A Simple Guide and Reference*. Allyn & Bacon.

Ghalib, T. K., & Al-Hattami, A. A. (2015). Holistic versus analytic evaluation of EFL writing: A case study. *English Language Teaching*, *8*(7), 225-236.

Guilford, J. P. (1973). *Fundamental Statistics in Psychology and Education*. McGraw-Hill.

Halaweh, M. (2023). ChatGPT in education: Strategies for responsible implementation. *Contemporary Educational Technology*, *15*(2).

Hong, W. C. H. (2023). The impact of ChatGPT on foreign language teaching and learning: Opportunities in education and research. *Journal of Educational Technology and Innovation*, *5*(1), 37-45.

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.

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... Hüllermeier, E. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*.

Khalil, M., & Er, E. (2023). Will ChatGPT get you caught? Rethinking of plagiarism detection. *arXiv*.

King, M. R. (2023). The future of AI in medicine: A perspective from a chatbot. *Annals of Biomedical Engineering*, *51*(2), 291-295.

Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. *RELC Journal*, *54*(2), 537-550.

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.

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., ... Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS Digital Health*, *2*(2).

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

Li, L. (2014). Experimental study on the validity of AES systems in the college EFL classroom. In *Proceedings of the 2nd International Conference on Teaching and Computational Science.*

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).

Moorhouse, B. L. (2023). Teachers' digital technology use after a period of online teaching. *ELT Journal*, *77*(4), 445-457.

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.

Perdana, I., & Farida, M. (2019). Online grammar checkers and their use for EFL writing. *Journal of English Teaching, Applied Linguistics and Literatures*, *2*(2), 67-76.

Rudner, L. M., Garcia, V., & Welch, C. (2006). An evaluation of IntelliMetric™ essay scoring system. *The Journal of Technology, Learning, and Assessment*, *4*(4), 1-22.

Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?. *Journal of Applied Learning and Teaching*, *6*(1), 1-22.

Sathyabalan, M., & Christian, M. (2022). A study of computer aided tools for evaluation of text in English as a foreign language. *2022 1st International Conference on Information System & Information Technology (ICISIT).*

Susnjak, T. (2022). ChatGPT: The end of online exam integrity?. *arXiv*.

Taecharungroj, V. (2023). What can ChatGPT do? Analyzing early reactions to the innovative AI chatbot on Twitter. *Big Data and Cognitive Computing*, *7*(1).

Tang, Z., & Cao, J. (2021). Language proficiency and syntactic complexity of Chinese EFL writers: A corpus-based study. *Forest Chemicals Review,* 1079-1090.

Tao, Y., & Yan, G. (2020). Characteristics of the use of cleft sentences in English majors' compositions. *Sino-US English Teaching*, *17*(2), 58-64.

Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research*, *10*(2), 157-180.

Wen, Q., Liang, M., & Yan, X. (2008). *Spoken and written corpus of Chinese learners*. Foreign Language Teaching and Research Press.

Wu, Z. Y. (2018). Can an automatic essay scoring system be used to improve students' writing skills. *International Journal of English Research*, *4*(2), 21-24.

Yan, H. (2019). I think we should…: Investigating lexical bundle use in the speech of English learners across proficiency levels. *International Journal of Translation, Interpretation, and Applied Linguistics*, *1*(2), 65-80.

Yan, H., & Li, Y. (2019). Beyond length: Investigating dependency distance across L2 modalities and proficiency levels. *Open Linguistics*, *5*(1), 601-614.

Yang, Y., Yap, N. T., & Mohamad Ali, A. (2022a). A corpus-based comparative study on syntactic complexity in university students' EFL writing in Southwestern China: A case of Pu'er University. *World Journal of English Language*, *12*(8), 172-180.

Yang, Y., Yap, N. T., & Mohamad Ali, A. (2022b). A review of syntactic complexity studies in context of EFL/ESL writing. *International Journal of Academic Research in Business and Social Sciences*, *12*(10), 441-454.

Yang, Y., Yap, N. T., & Mohamad Ali, A. (2023a). Chinese EFL learners' writing evaluation by ChatGPT. *Mendeley Data*.

Yang, Y., Yap, N. T., & Mohamad Ali, A. (2023b). Predicting EFL expository writing quality with measures of lexical richness. *Assessing Writing*, *57*.

Zhai, X. (2022). ChatGPT user experience: Implications for education. *SSRN*, 1-18.

Zhang, S. (2021). Review of automated writing evaluation systems. *Journal of China Computer-Assisted Language Learning*, *1*(1), 170-176.

Zimmerman, A. (2023). A ghostwriter for the masses: ChatGPT and the future of writing. *Annals of Surgical Oncology*, *30*, 3170-3173.

Yang, Y., & Zheng, Z. (2024). A refined and concise model of indices for quantitatively measuring lexical richness of Chinese university students' EFL writing. *Contemporary Educational Technology*, *16*(3).

Author Details

Yang Yang, *Southwest University, China, **Email ID**: yangvictoryang@swu.edu.cn*