Shabara, R., ElEbyary, K., & Boraie, D. (2024). TEACHERS or CHATGPT: The ISSUE of ACCURACY and CONSISTENCY in L2 ASSESSMENT. *Teaching English with Technology*, *24*(2), 71–92. https://doi.org/10.56297/vaca6841/LRDX3699/XSEZ5215

ISSUES

TEACHERS or CHATGPT: The ISSUE of ACCURACY and CONSISTENCY in L2 ASSESSMENT

Ramy Shabara[1], Khaled ElEbyary[2], Deena Boraie[3]

[1] Academic Services, University of Hertfordshire, hosted by The Global Academic Foundation, New Administrative Capital, Egypt, [2] Department of Education, University of York, UK and Damanhour University, Egypt, [3] TIRF (The International Research Foundation for English Language Education)

Keywords: ChatGPT, accuracy, consistency, intra-rater reliability, inter-rater reliability

https://doi.org/10.56297/vaca6841/LRDX3699/XSEZ5215

Teaching English with Technology

Vol. 24, Issue 2, 2024

Although there are claims that ChatGPT, an AI-based language model, is capable of assessing the writing of L2 learners accurately and consistently in the classroom, a number of recent studies have shown discrepancies between AI and human raters. Furthermore, there is a lack of studies investigating the intra-reliability of ChatGPT scores. Accordingly, this study aimed to examine the accuracy and consistency of ChatGPT compared to teachers, as well as with itself, after being trained on a rubric. To accomplish this goal, the study adopted a quantitative correlational non-experimental design. A dataset of 100 writing assignments, submitted by a cohort of B1-level students at an international branch university in Egypt, was analyzed quantitatively. These assignments were initially evaluated and moderated by trained teachers (n=11), and subsequently, the same assignments were also assessed twice by ChatGPT. The findings indicated that teachers' scores exhibited a higher level of accuracy compared to those generated by ChatGPT. The results also revealed that ChatGPT exhibits a moderate, yet questioned, level of intra-rater reliability. The weak-to-moderate correlations between ChatGPT and teacher scores raise concerns about the accuracy and consistency of ChatGPT's scoring of writing assignments. The implications of the findings highlight the potential applications and limitations of ChatGPT in L2 writing assessment. This study contributes to the ongoing discourse on the use of AI technologies in language education and provides insights into the accuracy and reliability of ChatGPT as an evaluation tool for L2 writing.

1. Introduction

Although AI research has been around since the 1950s, the emergence of ChatGPT has moved the goalposts closer to the extent that it has become essential to look at the impact of this tool more significantly and critically. ChatGPT has been identified as a chatbot that is capable of responding in a conversational manner to users' prompts that might require the system to do a range of tasks including commenting on something, writing compositions, summarizing literature, as well as a range of other tasks (Taecharungroj, 2023). However, there seems to be a continuum with tension between opponents and proponents of the use of ChatGPT in education. This has been described as an incited debate (Steiss et al., 2023). Opponents express serious ethical concerns about dehumanizing educational practices

and intellectual property and academic integrity issues in relation to incurring academic misconduct on the part of learners (Chomsky, 2023; Rudolph et al., 2023; Shoufan, 2023). Following this line of thought, some educational institutions banned the use of ChatGPT, such as New York Department of Education (Mohamed, 2023) and the University of Hong Kong (Yau & Chan, 2023), and commentators described this tool as a "plague" (Weissman, 2023).

On the other hand, some potential benefits of ChatGPT have attracted much attention (Grassini, 2023; Ray, 2023), and advocates believe that with the right regulations and ethics, ChatGPT (and similar tools) can positively be integrated into L2 education in the same way the internet and social media was when they came into existence. Hence, various researchers refused banning ChatGPT in schools and strongly supported the need to use it in teaching (Roose, 2023). Such polarity has been described by Garcia-Peñalvo (2023, p. 1) as "ranging from the enthusiasm of innovators and early adopters to the almost apocalyptic terror of the Terminator movie".

Therefore, the use of ChatGPT has caused both enthusiasm and dubiousness (Shoufan, 2023) because its possible influences on education are still largely unknown (Zhai, 2022). Indeed, the uncertainty has even been reflected in a number of publication titles such as "To resist it or to embrace it?" (Guo & Wang, 2024), "Is ChatGPT a blessing or a curse?" (Fuchs, 2023) and "ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?" (Rudolph et al., 2023). Nonetheless, L2 researchers describe ChatGPT as:

- the world's most advanced chatbot thus far (Liu & Ma, 2024; Rudolph et al., 2023),
- a technological innovation that has a hard-to-predict behaviour (García-Peñalvo, 2023),
- valuable, "cutting-edge", and hold[ing] considerable promise in revolutionizing EFL education (Koraishi, 2023; Ray, 2023),
- a potential chance to improve second language learning and instruction (Nguyen, 2023),
- can revolutionize the ways in which language is learned, taught, and assessed (Liu & Ma, 2023),
- may drive changes to educational learning goals, learning activities, and assessment and evaluation practices (Zhai, 2022).

Mohamed (2023) stated that ChatGPT has been used in several language learning scenarios such as "language tutoring, language generation, and language translation" (p. 3196).

Unlike claims in the literature that ChatGPT is capable of assessing and analyzing L2 learners' writing, the researchers involved in this study advocate a position of uncertainty regarding the ability of ChatGPT to assess L2 writing accurately and consistently in the classroom. They argue that the area is still largely unexplored. This study, therefore, examines the accuracy and consistency of the scoring behavior of ChatGPT and compares these with teachers' scoring behavior. The findings of the study aim to show the extent to which ChatGPT corresponds with or deviates from human assessors in assessing L2 writing, therefore, contributing valuable insights and advancing discussions to the ongoing discourse on the potential applications and limitations of ChatGPT in the assessment of L2 learners' writing.

2. Literature review

2.1. Accuracy and consistency in assessment

The concept of score accuracy pertains to the extent to which an assessment score reflects the true ability of a test taker on a measured construct (Kunnan, 2000). A score is considered accurate when it provides a reliable and valid representation of the construct being measured without being unduly influenced by construct-irrelevant factors such as bias, inconsistency, or measurement error (Xi, 2010). Highly accurate scores display compelling validity evidence — allowing defensible inferences about an examinee's status relative to defined criteria (Kane, 2013). Accuracy is crucial for meaningful assessment because inaccurate scores can unfairly disadvantage test-takers by misrepresenting their true proficiency levels, and, consequently, they can compromise the validity of test scores (Kunnan, 2004). Evaluating accuracy involves comparing scores to externally validated benchmarks or expert judgments (Bejar et al., 2004).

The consistency of assessment scores, also known as reliability, refers to the degree of stability and dependability of assessment scores across different administrations, forms, and scoring raters (Erford, 2013; Livingston, 2018). It is also concerned with the extent to which a score is free of error and the proportion of variation in scores (Bannigan & Watson, 2009). Consistency is essential because it does not only affect the interpretation and use of assessment scores but also it influences subsequent conclusions and decisions (American Educational Research Association et al., 2014; Bardhoshi & Erford, 2017).

In language testing, automated scoring is claimed to be equal or more accurate and consistent than human scoring (Wilson & Andrada, 2016). The main reason for this argument is the reliance of automated scoring systems on fixed scoring models which substantially differ from the subjective and error-prone nature of human evaluations (Zhang et al., 2020).

2.2. Previous studies on accuracy and consistency of AWE systems

Automated writing evaluation systems, exemplified by ChatGPT, offer L2 writers a simulated human-like automated scoring and feedback experience, designed to assess and improve their writing skills. Automated scoring refers to scores assigned by computers or artificial intelligence scoring technologies (Breyer et al., 2017; Ramesh & Sanampudi, 2021). Unlike automated feedback that offers guidance to writers with assessment-for-learning purposes in mind, automated scoring is mainly concerned with assessment-of-learning (Shi & Aryadoust, 2022). Although the two features are commonly integrated into AWE systems (Burstein et al., 2020), it is necessary to distinguish them because of the different purposes they serve and the validation methods and evidence required (Chapelle et al., 2015; Weigle, 2013).

Though they are widely used in high-stakes and classroom assessments, study results on the accuracy and consistency of AWE systems have been inconclusive (Deane, 2013b; Shi & Aryadoust, 2022, 2024) and ChatGPT is not an exception. In an early study, Burstein and Chodorow (1999) examined the scoring performance of an AWE system, namely e-rater, on a high-stakes test (i.e., the Test of Written English (TWE)) (n=510 essays). A comparison of e-rater scoring and human scoring means showed a statistically significant difference ($F(1,1128)= 5.469$, $p < .05$). In another study, conducted by Attali and Burstein (2006), a comparison between e-rater (Version 2) and human scoring on 25,000 essays revealed contrasting results, demonstrating a remarkably strong correlation of up to 0.97 between the machine and human raters. As a compromise, Bridgeman and colleagues (2012) extended the assessment of the validity of automated scoring beyond comparing it to human scores; they also considered predictive ability across different essays and times, as well as fairness toward subgroups. Analyzing large datasets from TOEFL and GRE essay responses segmented by gender, ethnicity, and country subgroups (n= 132,347), they found high overall agreement between human and machine scoring with correlations equal or sometimes higher between human-machine than between two human scorers. Tsai (2012) also investigated how closely human raters and an automated essay scoring system agree in their evaluation of high school students' English essays (n=923). The findings indicated that the agreement between human raters was significantly higher than the agreement between human raters and the automated essay scoring system. Similarly, Cohen et al. (2018) evaluated the reliability and validity of an automated essay scoring (AES) system compared to human raters. Two sets of 250 essays were assessed by both human raters and the AES system. The results indicated that the automated scores exhibited a comparable level of correlation with human scores, similar to the correlation observed among human raters, but there was a discrepancy between the reliability and validity of the AES; that is, although AES scoring was more consistent than human ratings, it was less valid. Similarly, Lu (2019) used

an AI writing evaluation system, Juku, to evaluate a set of 114 English essays written by Chinese college students. The study indicated that the system was less effective than human raters in providing accurate evaluations. Utilizing a GPT model by Mizumoto and Eguchi (2023) resulted in similar results. That is, the researchers evaluated the consistency and accuracy of the GPT-3 text-davinci-003 model in scoring essays from the ETS Corpus of Non-Native Written English (n=12,100) compared to human scoring using a 10-point rubric. The results indicated a certain level of consistency and accuracy in scoring though there was not perfect agreement with human raters. However, surprisingly, the authors suggested using AI tools, especially ChatGPT, as revolutionary and effective instruments in the AWE realm. Geçkin and colleagues (2023) reported similar findings when assessing the agreement between Chat-GPT 3.5 and five raters holistically scoring 43 paragraphs authored by a group of Turkish college students. Again, despite discrepancies, the authors highlighted that ChatGPT scores, paired with a human rater, can yield reliable and consistent results. In the same vein, Yancey et al. (2023) examined the accuracy of GPT-3.5 and GPT-4 in rating short essay responses composed by L2 English learners on a high-stakes language test (i.e., Duolingo) in comparison to human scores. Results suggested that, with calibration examples, GPT-4 approaches the performance of modern AWE methods, though agreement with human ratings may vary based on the test-taker's first language. In a review study, Shadiev and Feng (2023) reviewed 11 studies that examined the accuracy of various automated tools in evaluating texts and delivering feedback. The accuracy levels of these automated tools varied, with Pigai scoring at 45.5% and SpanishChecker at 94.5%. These inconsistent study findings bring up questions about the accuracy of AWE systems and the outcomes they produce.

Skeptical claims about the accuracy and consistency of AWE systems have been attributed to several reasons. One of the main concerns is susceptibility to manipulation. That is, AWE systems have the potential to be gamed, resulting in inflated scores compared to those given by human raters (Shi & Aryadoust, 2022). Additionally, critiques highlight the lack of direct connections between AWE system models and how human raters typically process and evaluate responses (Sari & Han, 2021; Shin & Gierl, 2021). The definition and representation of the writing construct are additional concerns (Condon, 2013; Roscoe et al., 2017; Shi & Aryadoust, 2022). Critics further cast doubt on the ability of AWE systems to assess highly cognitive-demanding skills such as audience awareness and critical thinking (Zhang, 2013) and to understand complex texts (Lu, 2019; Perin & Lauterbach, 2018). Moreover, opponents argue that AWE systems tend to prioritize surface-level language features over deeper ones, potentially overlooking crucial aspects of writing quality (Dikli, 2006; Shin & Gierl, 2021; Yun, 2023). These systems also fall short in recognizing the social nature of writing (Deane, 2013b; Woodworth & Barkaoui, 2020) and addressing intricate elements such as logic, clarity, accuracy, relevance, and

coherence, to name a few (Anson et al., 2013). Questions have also been raised about the transparency of the modeling process, the interpretability of scoring algorithms, and the overall accuracy of error detection and predictions (Zaidi, 2016). All these issues contribute to skepticism regarding the consistency and accuracy of AWE systems as well as consequent inferences and decisions (Vo et al., 2023).

2.3. What is missing in the literature?

Despite the growing interest in AWE systems and their potential benefits, research in this area has predominantly focused on specific dimensions. Vo and colleagues (2023) delineate three key areas that prior AWE research has tackled: (1) the alignment between automated and human-generated scores, assessed through reliability studies, (2) the fairness of AES ratings across different test-taker subgroups, investigated through analyses of ethnic group mean differences and differential item functioning (DIF), and (3) the associations of automated scores with external variables and the overall effect of AES use. However, there remains a scarcity of studies scrutinizing generative pre-trained transformer (GPT) technology, especially ChatGPT, in the context of AWE (Mizumoto & Eguchi, 2023). Also, to the best of the researchers' knowledge, there has been no investigation into ChatGPT's generated and regenerated scoring (i.e., intra-reliability) to date. The importance of delving into this issue stems from claims that ChatGPT is a game changer and could potentially substitute teachers in the AWE process (Geçkin et al., 2023). Furthermore, little research, if any, has investigated ChatGPT's rating scores in English for Academic Purposes programs involving undergraduate EFL students of different majors. Also, very few studies have trained ChatGPT on sample writings before vetting its consistency (e.g., Yancey et al., 2023). There seems to be a dearth of investigations into the accuracy of ChatGPT's scoring compared to itself and human raters as well.

3. Study

3.1. Research questions

The inconsistency of findings on the accuracy and consistency of AWE systems in scoring L2 writing, on the one hand, and the limited research on ChatGPT's scoring, on the other, result in knowledge void in the AWE domain, and consequently, limit our understanding of these issues. To address these gaps, the current work aims to investigate the primary research question: To what extent is ChatGPT accurate and consistent in assessing L2 writing compared to itself and teachers? This overarching question is dissected into the following sub-questions:

- RQ1. To what extent do ChatGPT-generated scores of written outputs differ from/agree with its own regenerated scores?

- RQ2. To what extent do ChatGPT-generated scores of written outputs differ from/agree with teacher-moderated scores?

3.2. Design

To address the research questions, this study adopted a quantitative correlational non-experimental design. The rationale for this design was to compare ChatGPT-(re)generated scores to those produced by itself and humans and investigate whether they are correlated.

3.3. Sample

The dataset for this study consisted of 100 cause-effect essays. These essays were selected randomly from a larger pool (n=599) and were written by EFL undergraduate students majoring in Pharmacy, Physiotherapy, Business, and Engineering and Computer Science within an international foundation program at an international branch university in Egypt. The L1 of all participants in this study was Arabic. The students were at the CEFR B1 English proficiency level and their ages ranged from 18 to 19 years old. They came from various high school backgrounds, including IGSCE, American Diploma, and public schools. Of the 100 essays, 52 were written by female students (52%). The essays were marked and moderated by a group of 11 trained and experienced teachers. Among these instructors, three (2.47%) were males. The teachers held various post-graduate degrees such as CELTA certifications (n=7) and MA degrees (n=4). Their professional years of experience ranged from 6 to 20 years with an average of 12.9 years (Median=10, SD=4.58).

3.4. Instrumentation

3.4.1. ESSAYS

The written essays constituted part of the students' assigned coursework in an EAP course. Students were given a one-week period to finalize their essays at home, using computers, and submit them. The essays were in response to a prompt asking students to write 350-450-words on the following topic:

> "Many students choose to seek part-time jobs during their studies." *What are the factors that contribute to students obtaining part-time jobs, and what are the effects of having a part-time job on students?*

To guide their writing, students were provided with specific sources that they were expected to integrate and cite within their essays.

3.4.2. RUBRIC

The analytic rubric used was developed by the parent university and has been effectively used for several years. All the academics in the branch campus received initial training from the parent university to ensure consistent application. In the branch university, the rubric has been in use for almost four years, with results consistently affirming its consistency and accuracy.

The rubric consisted of three criteria: (1) Organization, Content and Relevance (O, C & R); (2) Language Use (LU); and (3) Communicative Quality, Use of Academic Vocabulary and Style (CQ, UoAV & S). Each criterion was individually rated on a scale of 100 with specific performance descriptors of 10 bands (0-100). The final score for each essay was the average of the three scores, reflecting performance across all criteria.

3.4.3. AUTOMATED WRITING EVALUATION SYSTEM

The AWE system used in the study was ChatGPT 3.5. It was used to give analytic and holistic scores to the written essays under investigation based on the above-mentioned rubric.

3.5. Procedures

To conduct the study, ethical clearance was first obtained from the Ethics Committee with Delegated Authority (ECDA) under the Protocol Number SLE/SF/UH/06071 of an international university where one of the researchers was affiliated. The approval gave permission to conduct the study and access the Learning Management System (LMS) for data collection.

A random sample of 100 human-marked and moderated essays was extracted from the university's LMS to be scored by ChatGPT. Each essay had initially been marked by a well-trained and experienced teacher who was familiar with an analytic rubric employed for almost four years. The marking process was further validated through moderation by another trained and experienced teacher. In the events of discrepancies, a third teacher, typically a module leader, was consulted to ensure consistency in the marking process. In all cases, revised scores were assigned only after the three raters reached a consensus to validate the scores and ensure their consistency.

To undertake the marking process with ChatGPT, three phases were followed. Firstly, a conversation with ChatGPT was run where it was asked about its potential to mark university essays based on an analytic rubric. Secondly, based on its affirmation, ChatGPT was trained to use the rubric. It was provided with the rubric and seven samples of written assignments from seven different bands of the rubric, specifically the 30s-90s as almost all essays generally fell within these bands. This training was essential for two main reasons: to avoid ChatGPT's misuse of the rubric, and to ensure it understood and followed the rubric used by teachers. The same training

was repeated one more time and the seven essays were provided to ChatGPT in a different order to make sure that Chat GPT followed the rubric and assigned accurate scores. Once it was confirmed that ChatGPT adhered to the rubric in scoring with appropriate justification, the 100 essays were coded and assessed by ChatGPT, based on the same rubric utilized previously by teachers. To guide the scoring process, ChatGPT was provided with a specific prompt. The prompt instructed ChatGPT to imagine itself as an English for Academic Purposes teacher at the university level and assess a student's essay on the topic of part-time jobs during studying. The prompt also presented the specific criteria, performance bands, and descriptors for evaluation. Based on the provided rubric, ChatGPT scored each essay analytically, and then holistically by averaging the analytic scores. Following the initial scoring process, the 100 essays were reshuffled randomly and presented to ChatGPT for a second scoring process. To ensure complete independence of essay evaluations, a new GPT conversation was initiated for each essay to be evaluated. The scores obtained from ChatGPT in the two scoring processes were compared to assess its own intra-rater reliability, while the first scores of ChatGPT were compared to the moderated scores of teachers to vet its inter-rater reliability.

3.6. Analysis

To assess the accuracy and consistency of ChatGPT scoring, two statistical techniques were used, namely descriptive and inferential. Descriptive statistics (i.e., means and standard deviations) were employed to compare the scores provided by ChatGPT and those of teachers, as the benchmark. For a more nuanced examination, inferential statistics, encompassing a t-test (paired-samples and independent) and intraclass correlation coefficient (ICC) (a two-way, random effect model) were computed. The rationale for incorporating t-tests alongside ICC lies in the recognition that high correlations between AES scores and human ratings alone do not suffice to validate AES score use (Barkaoui & Woodworth, 2023) because agreement results alone do not fully capture the construct measured by automated scores (Attali, 2007).

The paired-samples and independent t-tests were manipulated to measure the statistical significance of the difference between the scores assigned by ChatGPT and itself and ChatGPT and teachers, respectively, while ICC was employed to examine the consistency and agreement between ChatGPT's evaluations and both itself and teachers. ICCs are preferred over Pearson coefficients for measuring "the relationship between different measures of the same construct [as it] does account for individual variability" (Cole et al., 2013, p. 733). Additionally, it considers both the correlation between raters and the differences between the actual scores assigned (Larsen-Hall, 2010).

4. Results

4.1. To what extent do ChatGPT-generated scores of written outputs differ from/agree with its own regenerated scores?

Table 1 presents the mean scores, standard deviations, intraclass correlation coefficients (ICC) for ChatGPT, and the results of a paired-samples t-test conducted to compare ChatGPT's scores and its own performance.

Table 1. Mean scores, standard deviations, intraclass correlation coefficients & paired-samples t-test for ChatGPT scores across two occasions

	Analytic scores						Holistic scores	
Criteria	O, C & R		LU		CQ, UoAV & S			
ChatGPT	Ch1	Ch2	Ch1	Ch2	Ch1	Ch2	Ch1	Ch2
M	61.47	59.35	53.67	51.51	55.32	52.66	56.70	54.62
Mean difference	2.12		2.16		2.66		2.08	
SD	9.14	9.36	11.15	11.82	12.60	11.21	10.72	10.04
ICC	.69**		.58**		.66**		.69**	
ICC 95% CI	L= 0.54	U=0.79	L=0.37	U=0.71	L=0.50	U=0.77	L=0.54	U=0.79
t(df=99)	2.4		1.7		2.2		2.1	
P-value	.019***		.087		.027***		.041***	

**ICC is significant at the 0.01 level.

***Paired-samples t-test is significant at the 0.001 level.

Notes: CI= Confidence Interval L= Lower U=Upper Values are rounded up.

Analytically, the intra-rater reliability, indexed by ICCs, were generally moderate/fair (Koo & Li, 2016). That is, ChatGPT's scores showed moderate agreement with its own regenerated scores for Organization, Content and Relevance (r_{ICC}= .69, p < .01, 95% CI [.54—.79]), Language Use (r_{ICC} = .58, p < .01, 95% CI [.37—.71]), and Communicative Quality, Use of Academic Vocabulary and Style (r_{ICC} = .66, p < .01, 95% CI [.50—.77]). The mean scores assigned by ChatGPT (Ch1) for the three criteria were slightly higher compared to the regenerated scores (Ch2) (M= 61.46 vs. 59.35 for *Organization, Content and Relevance*; 53.67 vs. 51.51 for *Language Use*; and 55.32 vs. 52.66 for *Communicative Quality, Use of Academic Vocabulary and Style*). Standard deviations were almost similar between Ch1 and Ch2 for Organization, Content and Relevance (SD= 9.14 vs. 9.36), Language Use, (SD=11.15 and 11.82), and Communicative Quality, Use of Academic Vocabulary and Style (SD= 12.60 vs. 11.21), respectively.

For the holistic scores, there was a moderate correlation between ChatGPT's scores and its regenerated scores (r_{ICC} = .69, p < .01, 95% CI [.54-.79]). The mean holistic scores were close between Ch1 (M= 56.70) and Ch2 (M= 54.62). Standard deviations were comparable between Ch1 and Ch2 holistic scores (SD= 10.72 vs. 10.04), respectively.

Overall, the results suggest that ChatGPT exhibits a moderate, yet questioned, level of intra-rater reliability. The mean scores being slightly higher in the first scoring (Ch1) versus the second (Ch2) could indicate a practice effect, with ChatGPT's scoring becoming slightly harsher on the second try. But overall, the means are fairly close.

In terms of significant differences between the mean scores of ChatGPT and itself, there was a statistically significant difference between the mean scores generated and regenerated by ChatGPT analytically for the tests for Organization, Content and Relevance ($t= 2.4$, $df= 99$, $p < 0.001$), and Communicative Quality, Use of Academic Vocabulary and Style ($t= 2.2$, $df= 99$, $p < 0.001$), as well as holistically ($t= 2.1$, $df= 99$, $p < 0.001$). However, there was no statistically significant difference between the mean scores for the Language Use criterion ($t= 0.99$, $df= 385$, $p= 0.087$, two-tailed). The statistically significant differences between ChatGPT-generated and regenerated mean scores give rise to three concerns about score accuracy, consistency and the way ChatGPT uses the rubric for assigning scores.

4.2. To what extent do ChatGPT-generated scores of written outputs differ from/agree with teacher-moderated scores?

Descriptive and inferential statistics (i.e., mean scores and standard deviations, intraclass correlation coefficients (ICC), independent-samples t-test, and effect size (d)) of ChatGPT's and teachers' scores are displayed in Table 2.

Table 2. Mean scores, standard deviations, intraclass correlation coefficients, independent-samples t-test & effect sizes for ChatGPT's and teachers' scores

	Analytic scores						Holistic scores	
Criteria	O, C & R		LU		CQ, UoAV & S			
ChatGPT (Ch1) vs. Teachers (T)	Ch1	T	Ch1	T	Ch1	T	Ch1	T
M	61.47	48.24	53.67	46.90	55.32	46.98	56.70	47.55
Mean difference	13.23		6.77		8.34		9.15	
SD	9.14	9.80	11.15	10.00	12.60	10.50	10.72	9.66
ICC	.30**		.43**		.57**		.47**	
ICC 95% CI	L=-0.17	U=0.59	L=0.12	U=0.63	L=0.16	U=0.76	L=0.00	U=0.70
t(df=198)	9.87		4.52		5.08		6.34	
P-value	.000***		.000***		.000***		.000***	
Effect size (Cohen's d)	1.4 (Large)		0.64 (Medium)		0.72 (Medium)		0.9 (Large)	

**ICC is significant at the 0.01 level.

***Independent-samples t-test is significant at the 0.001level.

Notes: CI= Confidence Interval L= Lower U=Upper Values are rounded up.

Results revealed that, analytically, there was a weak/ poor inter-rater reliability, indexed by ICCs (Koo & Li, 2016), between ChatGPT's scores and teacher-moderated scores for Organization, Content and Relevance ($r_{ICC}= .30$, $p < .01$, 95% CI [-.17—.59]). Notably, ChatGPT's mean scores

(M= 61.47, SD= 9.14) were considerably higher than those of the teachers (M= 48.24, SD= 9.80). Similarly, in terms of the Language Use criterion, the agreement was weak (r_{ICC}= .42, p < .01, 95% CI [.12—.63]), and once again, ChatGPT's mean scores (M= 53.67, SD= 11.15) surpassed the teachers' scores (M= 46.90, SD=10.00). Conversely, for Communicative Quality, Use of Academic Vocabulary, and Style, findings indicated a moderate/ fair agreement between ChatGPT's and teachers' scores (r_{ICC}= .57, p < .01, 95% CI [.16—.76]). Remarkably, ChatGPT's mean scores (M= 55.32, SD= 12.60) remained higher than those of the teachers (M= 46.98, SD= 10.50).

Holistically, there was a weak inter-rater reliability between ChatGPT's and teachers' moderated scores (r_{ICC} = .47, p < .01, 95% CI [.00—.70]). ChatGPT's mean holistic scores (M= 56.70, SD= 10.72) were about 10 points higher than the teachers' (M= 47.55, SD= 9.66).

Overall, reliability and accuracy concerns arise from the observed weak positive correlations. In terms of reliability, the considerably higher mean scores consistently assigned by ChatGPT in comparison to teachers both analytically, across all criteria, and holistically raise further concerns about the consistency of ChatGPT scoring, implying a consistent leniency in its scoring approach. This discrepancy is further emphasized by the higher standard deviations in ChatGPT's scores, suggesting a more variable scoring distribution compared to teachers, thereby compromising overall score reliability. Similarly, the low correlations between ChatGPT and teacher scores raise doubts about the accuracy of ChatGPT's scoring. That is, its scores do not strongly align with the teachers' moderated scores. The higher variability in ChatGPT's scores, embodied by mean scores and standard deviations, also suggests potential issues with construct validity in terms of measuring the underlying attributes such as organization and language use in a reliable way.

Regarding the significant discrepancies in mean scores between ChatGPT and human teachers, noteworthy dissimilarities emerged both analytically and holistically. Analytically, substantial differences were observed across the criteria of Organization, Content, and Relevance (M= 61.47, SD= 9.14 vs. M= 48.24, SD = 9.80; $t(198)$= 9.87, p< 0.001), Language Use (M= 53.67, SD= 11.15 vs. M= 46.90, SD= 9.80; $t(198)$= 4.52, p< 0.001), and Communicative Quality, Use of Academic Vocabulary, and Style (M=55.32, SD= 12.60 vs. M= 46.98, SD = 10.50; $t(198)$= 5.08, p< 0.001). The effect size for the difference in mean scores for Organization, Content, and Relevance was considerably large (d= 1.4), while for the other two criteria, it was within the medium range (d= 0.64 and 0.72, respectively).

Holistically, a statistically significant difference occurred between ChatGPT-generated scores (M=56.70, SD= 10.72) and moderated human-assigned scores (M= 47.55, SD= 9.66; $t(198)$= 6.34, p < 0.001), accompanied by a

substantial Cohen's d value (0.9). These significant differences, coupled with medium-large effect sizes, raise further concerns regarding the accuracy of scores assigned by ChatGPT compared to human teachers.

5. Discussion

This study scrutinized the consistency and accuracy of ChatGPT's scoring compared to itself and teachers. To achieve this goal, 100 L2 essays were analytically and holistically scored by ChatGPT on two occasions, and then, they were juxtaposed to teacher-moderated scoring. Overall findings imply limitations in ChatGPT's current scoring capabilities.

To elaborate, the findings of Research Question 1 indicate that ChatGPT demonstrated moderate intra-rater reliability, with ICCs ranging from .58 (Language Use), .66 (Communicative Quality, Use of Academic Vocabulary and Style), to .69 (Organization, Content and Relevance). Additionally, significant score differences emerged between the original and repeated scoring with strict marking during the repeated trial. One possible justification of the moderate intra-rater reliability could be ChatGPT's algorithm whereby scoring is processed. It is acknowledged that ChatGPT's algorithm is obscure and the way it is manipulated to analyze inputs is not fully understood (Geçkin et al., 2023; Uto, 2021). Additionally, AWE systems tend to conceptualize and operationalize the construct of writing differently compared to human raters (Barkaoui & Woodworth, 2023). This is what may justify the shortfalls associated with AWE systems in terms of evaluating writing (Hussein et al., 2019) though modern AWE systems incorporate machine learning algorithms and deep learning-based approaches to evaluate writing quality (Chen & Pan, 2022). Another reason for the observed modest scoring consistency between ChatGPT and itself could be its non-deterministic characteristics where "identical input can lead to different outputs" (Reiss, 2023, p. 3), and consequently, affect scoring consistency. A third reason is that ChatGPT is not designed for this function. That is, ChatGPT is developed as a language generation model, and thus, it does not seem to be able to handle AWE functions (Escalante et al., 2023; Mizumoto & Eguchi, 2023). Another plausible factor contributing to the modest intra-rater consistency may be the rubric used in the study. It is possible that the 100-point rubric was difficult to be interpreted and consistently applied by ChatGPT. This, consequently, reinforces the need for more transparency of the AWE modeling processes and the interpretability of scoring algorithms (Zaidi, 2016).

The fact that ChatGPT produced lower scores in the second scoring process could be due to a potential practice effect which may be an inherent variability in AWE algorithms and the data on which it was trained (Wilson & Andrada, 2016). Interestingly, though research has yet to explore the real reasons for such a discrepancy, this drift in scoring is similar to human rating performance where raters may become more lenient or stricter over

time because of such factors as fatigue, exposure and various interpretations of scoring rubrics (Eckes, 2008). Accordingly, the findings imply that ChatGPT's scoring may be prone to similar biases. This emphasizes the need for continuous training, calibration, and monitoring to ensure consistent scoring performance.

In response to Research Question 2, the inter-rater reliability between ChatGPT and teachers, as measured by ICCs, is found to be mixed, with a moderate level of agreement for the Communicative Quality, Academic Vocabulary and Style criterion (r_{ICC}=.57) but weak agreement for the Organization, Content and Relevance, and Language Use criteria (r_{ICC}=.30 and .40, respectively) as well as holistic scoring, overall. Surprisingly, the lowest agreement, contrary to the intra-rater results, was for the Organization, Content and Relevance criterion. Furthermore, significant disparities in scoring were observed between ChatGPT and teacher scoring, except for the Language Use criterion, with a more lenient marking approach from ChatGPT. Taken together, the results raise serious concerns regarding the consistency and accuracy of ChatGPT evaluations of the quality of students' writing.

The discrepancy of correlation between ChatGPT and human scores (i.e., gold standards) can be attributed to the difference between how AWE systems and human raters process and define the writing construct. According to Wilson and Roscoe (2020), AWE systems reduce the construct of writing to merely assessing text length, syntax, and vocabulary, and they ignore the intricate sociocultural dynamics involved in writing. Moreover, AWE systems are sample-dependent (Zhang et al., 2020); that is, they rely on training with a corpus of human-scored essays that represent true scores (Correnti et al., 2019) – a feature notably absent in ChatGPT's design for scoring. Additionally, Barkaoui and Woodworth (2023) claimed that AES systems seem to use various criteria and/or assign different weights to the same criteria in a different way compared to human raters. This result underscores the importance of revisiting the algorithms of AWE systems, in general, and ChatGPT, in particular, to make sure that the scoring process and the writing construct are processed, defined, and operationalized in a similar way to that of human raters.

Previous research can explain the weak agreements between ChatGPT's and human raters' scores on the Organization, Content and Relevance criteria. ChatGPT has been reported to be inaccurate at assessing the organizational quality of essays because it is very sensitive to and programmed by the presence and absence of cohesive devices regardless of their appropriateness and it depends on superficial linguistic features to assess organizational quality (Yoon et al., 2023). Previous research has also highlighted the

challenges faced by AWE systems in accurately detecting higher-order writing skills such as organization (e.g., Barkaoui & Woodworth, 2023; Deane, 2013a; Dikli & Bleyle, 2014; Gardner et al., 2020; Vojak et al., 2011).

As AWE systems pay more attention to grammatical structures and word count over meaning (Wang & Brown, 2007), the inconsistent agreement between ChatGPT's and teachers' scores assigned to the Language Use and Communicative Quality, Academic Vocabulary and Style criteria led to inflated scores in favor of ChatGPT (Shi & Aryadoust, 2022). According to Vo and colleagues (2023), "By simply giving more weight to word count, AES can easily achieve a high level of correlation with human raters, without necessarily measuring the same construct as human raters do." (p. 2)

Collectively, the observation that intra-rater consistency indices are higher than inter-rater ones is well documented in the literature. This is aligned with Dikli and Bleyle's (2014) study in which they contended that "automated scoring systems tend to have higher internal consistency than human raters due to algorithmic scoring based on training data rather than more subjective human judgment". (p. 41) The weak correlation between AWE and human scoring, together with ChatGPT's leniency in scoring, aligns with the results of Huang's (2014) study. Huang concluded that automated essay scoring systems tend to assign higher marks than human scorers.

The overall disparities between ChatGPT's and teachers' scores can also be justified for several reasons. First, ChatGPT, like other AWE systems, seems to miss or miscode some errors (Woodworth & Barkaoui, 2020). It also appears that it overvalues or undervalues some writing aspects (Barkaoui & Woodworth, 2023). Previous studies have also found that some automated essay scoring systems exhibit variable scoring patterns over time (e.g., Zupanc & Bosnić, 2017). Taken together, this finding raises concerns about the consistency and accuracy of ChatGPT's scores, casting doubt on the feasibility of supporting or replacing teachers' evaluations with AI scoring systems. Accordingly, this stance contradicts the studies calling for employing ChatGPT as a revolutionary AWE system (e.g., Mizumoto & Eguchi, 2023; Shi & Aryadoust, 2024). It also lends support to the questions over the validity of the explanation of inference of assessments in AWE systems (Barkaoui & Woodworth, 2023). Consequently, these results carry substantial implications for the conclusions drawn about test-takers' abilities.

6. Conclusions and recommendations

In light of the above results, several conclusions and questions arise regarding the use of ChatGPT 3.5 as a reliable AWE tool to assess student essays. The results of this study showed that ChatGPT is not sufficiently consistent nor accurate in assessing writing quality and accordingly, it is not yet viable to substitute teachers. ChatGPT can support teachers by providing feedback to students on their writing, and consequently, it can save time for them.

Contrary to the claims that ChatGPT does not need to be trained on human corpora tailored for a particular task or genre (Steiss et al., 2024), more training on high-quality data and processing the intricate construct of writing is strongly recommended for ChatGPT to enhance its scoring performance.

Equally important, educators should address the limitations of AI tools and prioritize the development of AI literacy among students and teachers if they aim to use them effectively in writing instruction (Tate et al., 2023). The algorithm of ChatGPT needs to be revisited to be able to be used in the writing assessment process.

The limitations of the study also need to be addressed. First, using the free version of ChatGPT (v. 3.5), with its existing resources and pre-2021 data (Xiao & Zhi, 2023), may have limited access to the latest features or updates. Second, the complexity of the scoring rubric may have affected ChatGPT's scoring accuracy. The small number of essays used to train ChatGPT on the rubric poses another limitation for the study where it may have affected the consistency and accuracy of ChatGPT's scores.

Accordingly, similar studies can be conducted with a shorter and simpler rubric as well as a larger pool of writing samples to train ChatGPT. Further studies are also recommended to continue to explore the consistency and accuracy of various AI AWE tools to evaluate the most reliable and accurate system(s) compared to human raters. Moreover, further research is needed to address questions such as "What thresholds would be needed to assure that ChatGPT has human-level scoring accuracy and consistency?", "How can the correlations between ChatGPT's and human's scores be improved?", and "What measures other than correlations can be employed to enhance our understating of what is measured by AWE systems?".

Submitted: July 03, 2024 EEST

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association.

Anson, C., Filkins, S., Hicks, T., O'Neill, P., Pierce, K. M., & Winn, M. (2013). *NCTE Position Statement on Machine Scoring: Machine Scoring Fails the Test*. National Council of Teachers of English. http://www.ncte.org

Attali, Y. (2007). Construct validity of E-rater® in scoring TOEFL® essays. *ETS Research Report Series*, *2007*, i–220. https://doi.org/10.1002/j.2333-8504.2007.tb02063.x

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater®; V. 2. *The Journal of Technology, Learning and Assessment*, *4*(3). https://ejournals.bc.edu/index.php/jtla/article/view/1650

Bannigan, K., & Watson, R. (2009). Reliability and validity in a nutshell. *Journal of Clinical Nursing*, *18*, 3237–3243. https://doi.org/10.1111/j.1365-2702.2009.02939.x

Bardhoshi, G., & Erford, B. T. (2017). Processes and procedures for estimating score reliability and precision. *Measurement and Evaluation in Counseling and Development*, *50*(4), 256–263. https://doi.org/10.1080/07481756.2017.1388680

Barkaoui, K., & Woodworth, J. (2023). An exploratory study of the construct measured by automated writing scores across task types and test occasions. *Studies in Language Assessment*, *12*(1), 1–38. https://doi.org/10.58379/QCFS2805

Breyer, F. J., Rupp, A. A., & Bridgeman, B. (2017). Implementing a contributory scoring approach for the GRE® analytical writing action: A comprehensive empirical investigation. *ETS Research Report Series*, *2017*(1), 1–28. https://doi.org/10.1002/ets2.12142

Bridgeman, B., Trapani, C., & Yigal, A. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education*, *25*(1), 27–40. https://doi.org/10.1080/08957347.2012.635502

Burstein, J., & Chodorow, M. (1999). Automated essay scoring for nonnative English speakers. In *Proceedings of a Symposium on Computer Mediated Language Assessment and Evaluation in Natural Language Processing* (pp. 68–75). Association for Computational Linguistics. https://doi.org/10.3115/1598834.1598847

Burstein, J., Riordan, B., & McCaffrey, D. (2020). Expanding automated writing evaluation. In D. Yan, A. A. Rupp, & P. Foltz (Eds.), *Handbook of Automated Scoring: Theory into Practice* (pp. 329–346). Taylor and Francis Group/CRC Press. https://doi.org/10.1201/9781351264808

Chapelle, C. A., Cotos, E., & Lee, J. (2015). Validity arguments for diagnostic assessment using automated writing evaluation. *Language Testing*, *32*(3), 385–405. https://doi.org/10.1177/0265532214565386

Chen, H., & Pan, J. (2022). Computer or human: A comparative study of automated evaluation scoring and instructors' feedback on Chinese college students' English writing. *Asian-Pacific Journal of Second and Foreign Language Education*, *7*(34), 1–20. https://doi.org/10.1186/s40862-022-00171-4

Chomsky, N. (2023, March 8). The false promise of ChatGPT. *The New York Times*. https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html

Cohen, Y., Levia, E., & Ben-Simon, A. (2018). Validating human and automated scoring of essays against "True" scores. *Applied Measurement in Education*, *31*(3), 241–250. https://doi.org/10.1080/08957347.2018.1464450

Cole, W. R., Arrieux, J. P., Schwab, K., Ivins, B. J., Qashu, F. M., & Lewis, S. C. (2013). Test–retest reliability of four computerized neurocognitive assessment tools in an active duty military population. *Archives of Clinical Neuropsychology*, *28*, 732–742. https://doi.org/10.1093/arclin/act040

Condon, W. (2013). Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings? *Assessing Writing*, *18*, 100–108. https://doi.org/10.1016/j.asw.2012.11.001

Correnti, R., Matsumura, L. C., Wang, E., Litman, D., Rahimi, Z., & Kisa, Z. (2019). Automated scoring of students' use of text evidence in writing. *Reading Research Quarterly*, *55*(3), 493–520. https://doi.org/10.1002/rrq.281

Deane, P. (2013a). Covering the construct: An approach to automated essay scoring motivated by a socio-cognitive framework for defining literacy. In M. D. Shermis & J. Burstein (Eds.), *Handbook of Automated Essay Evaluation: Current Applications and New Directions* (pp. 298–312). Routledge.

Deane, P. (2013b). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, *18*, 7–24. https://doi.org/10.1016/j.asw.2012.10.002

Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment*, *5*(1), 1–36. https://ejournals.bc.edu/index.php/jtla/article/view/1640/1489

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

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*(2), 155–185. https://doi.org/10.1177/0265532207086780

Erford, B. T. (2013). *Assessment for Counselors* (2nd ed.). Cengage Wadsworth.

Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education*, *20*(57), 1–20. https://doi.org/10.1186/s41239-023-00425-2

Fuchs, K. (2023). Exploring the opportunities and challenges of NLP models in higher education: Is Chat GPT a blessing or a curse? *Frontiers in Education*. https://doi.org/10.3389/feduc.2023.1166682

García-Peñalvo, F. J. (2023). The perception of artificial intelligence in educational contexts after the launch of ChatGPT: Disruption or panic? *Education in the Knowledge Society*, *24*, 1–9. http://repositorio.grial.eu/handle/grial/2838

Gardner, J., O'Leary, M., & Yuan, L. (2020). Artificial intelligence in educational assessment: 'Breakthrough? Or buncombe and ballyhoo?' *Journal of Computer Assisted Learning*, *37*, 1207–1216. https://doi.org/10.1111/jcal.12577

Geçkin, V., Kızıltaş, E., & Çınar, Ç. (2023). Assessing second-language academic writing: AI vs. human raters. *Journal of Educational Technology & Online Learning*, *6*(4), 1096–1108. https://doi.org/10.31681/jetol.1336599

Grassini, S. (2023). Shaping the future of education: Exploring the potential and consequences of AI and ChatGPT in Educational Settings. *Education Sciences*, *13*(7), 692. https://doi.org/10.3390/educsci13070692

Guo, K., & Wang, D. (2024). To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. *Education and Information Technologies*, *29*, 8435–8463. https://doi.org/10.1007/s10639-023-12146-0

Huang, S. (2014). Automated versus human scoring: A case study in an EFL context. *Electronic Journal of Foreign Language Teaching*, *11*(1), 149–164.

Hussein, M. A., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: A literature review. *Peer Journal of Computer Science*, *5*(2), e208. https://doi.org/10.7717/peerj-cs.208

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Koraishi, O. (2023). Teaching English in the age of AI: Embracing ChatGPT to optimize EFL materials and assessment. *Language Education and Technology*, *3*(1), 55–72.

Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (Ed.), *Fairness and Validation in Language Assessment* (pp. 1–14). Cambridge University Press.

Kunnan, A. J. (2004). Test fairness. In M. Milanovic & C. Weir (Eds.), *Studies in Language Testing 18: European Language Testing in a Global Context* (pp. 109–132). Cambridge University Press.

Larsen-Hall, J. (2010). *A Guide to Doing Statistics in Second Language Research Using SPSS* (1st ed.). Routledge.

Liu, G., & Ma, C. (2024). Measuring EFL learners' use of ChatGPT in informal digital learning of English based on the technology acceptance model. *Innovation in Language Learning and Teaching*, *18*(2). https://doi.org/10.1080/17501229.2023.2240316

Livingston, S. A. (2018). *Test Reliability—Basic Concepts*. Educational Testing Service (ETS).

Lu, X. (2019). An empirical study on the artificial intelligence writing evaluation system in China CET. *Big Data*, *7*, 121–129. https://doi.org/10.1089/big.2018.0151

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, *2*, 1–13. https://doi.org/10.1016/j.rmal.2023.100050

Mohamed, A. M. (2023). Exploring the potential of an AI-based chatbot (ChatGPT) in enhancing English as a foreign language (EFL) teaching: Perceptions of EFL faculty members. *Education and Information Technologies*. https://doi.org/10.1007/s10639-023-11917-z

Nguyen, T. (2023). EFL teachers' perspectives toward the use of ChatGPT in writing classes: A case study at Van Lang University. *International Journal of Language Instruction*, *2*(3), 1–47. https://doi.org/10.54855/ijli.23231

Perin, D., & Lauterbach, M. (2018). Assessing text-based writing of low-skilled college students. *International Journal of Artificial Intelligence in Education*, *28*, 56–78. https://doi.org/10.1007/s40593-016-0122-z

Ramesh, D., & Sanampudi, S. K. (2021). An automated essay scoring system: A systematic literature review. *Artificial Intelligence Review*, *55*, 2495–2527. https://doi.org/10.1007/s10462-021-10068-2

Ray, P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, *3*, 121–154. https://doi.org/10.1016/j.iotcps.2023.04.003

Reiss, M. (2023). *Testing the Reliability of ChatGPT for Text Annotation and Classification: A Cautionary Remark*. https://doi.org/10.48550/arXiv.2304.11085

Roose, K. (2023, January 12). Don't ban ChatGPT in schools. Teach with it. https://www.nytimes.com/2023/01/12/technology/chatgpt-schools-teachers.html

Roscoe, R. D., Wilson, J., Johnson, A. C., & Mayra, C. R. (2017). Presentation, expectations, and experience: Sources of student perceptions of automated writing evaluation. *Computers in Human Behavior*, *70*, 207–2021. https://doi.org/10.1016/j.chb.2016.12.076

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. https://doi.org/10.37074/jalt.2023.6.1.9

Sari, E., & Han, T. (2021). Automated L2 writing performance assessment: A literature review. *The Reading Matrix: An International Online Journal*, *21*(2), 66–84. https://www.readingmatrix.com/files/25-7xdxc1r5.pdf

Shadiev, R., & Feng, Y. (2023). Using automated corrective feedback tools in language learning: A review study. *Interactive Learning Environments*, 1–29. https://doi.org/10.1080/10494820.2022.2153145

Shi, H., & Aryadoust, V. (2022). A systematic review of automated writing evaluation systems. *Education and Information Technologies*, *28*. https://doi.org/10.1007/s10639-022-11200-7

Shi, H., & Aryadoust, V. (2024). A systematic review of AI-based automated written feedback research. *ReCALL*, *36*(2), 187–209. https://doi.org/10.1017/S0958344023000265

Shin, J., & Gierl, M. J. (2021). More efficient processes for creating automated essay scoring frameworks: A demonstration of two algorithms. *Language Testing*, *38*(2), 247–272. https://doi.org/10.1177/0265532220937830

Shoufan, A. (2023). Exploring students' perceptions of ChatGPT: Thematic analysis and follow-Up survey. *IEEE Access*, *11*, 38805–38818. https://doi.org/10.1109/ACCESS.2023.3268224

Steiss, J., Tata, T., & Warschauer, M. (2023). *Emergent AI-assisted Discourse: Case Study of a Second Language Writer Authoring with ChatGPT*. https://arxiv.org/abs/2310.10903

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction*, *91*, 1–15. https://doi.org/10.1016/j.learninstruc.2024.101894

Taecharungroj, V. (2023). "What can ChatGPT do?" Analyzing early reactions to the innovative AI chatbot on twitter. *Big Data and Cognitive Computing*, *7*(1). https://doi.org/10.3390/bdcc7010035

Tate, T., Doroudi, S., Ritchie, D., Xu, Y., & Warschauer, M. (2023). *Educational Research and AI-generated Writing: Confronting the Coming Tsunami*. https://doi.org/10.35542/osf.io/4mec3

Tsai, M. (2012). The consistency between human raters and an automated essay scoring system in grading high school students' English writing. *Action in Teacher Education*, *34*, 328–335. https://doi.org/10.1080/01626620.2012.717033

Uto, M. (2021). A review of deep-neural automated essay scoring models. *Behaviormetrika*, *48*, 459–484. https://doi.org/10.1007/s41237-021-00142-y

Vo, Y., Rickels, H., Welch, C., & Dunbar, S. (2023). Human scoring versus automated scoring for English learners in a statewide evidence-based writing assessment. *Assessing Writing*, *56*, 1–16. https://doi.org/10.1016/j.asw.2023.100719

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

Wang, J., & Brown, M. S. (2007). Automated essay scoring versus human scoring: A comparative study. *Journal of Technology, Learning, and Assessment*, *6*(2), 4–28.

Weigle, S. C. (2013). English as a second language writing and automated essay evaluation. In M. D. Shermis & J. Burstein (Eds.), *Handbook of Automated Essay Evaluation: Current Applications and New Directions* (pp. 36–54). Routledge. https://doi.org/10.4324/9780203122761

Weissman, J. (2023). ChatGPT is a plague upon education. https://www.insidehighered.com/views/2023/02/09/chatgpt-plague-upon-education-opinion

Wilson, J., & Andrada, N. G. (2016). Using automated feedback to improve writing quality: Opportunities and challenges. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *The Handbook of Research on Technology Tools for Real-World Skill Development* (pp. 678–703). IGI Global. https://doi.org/10.4018/978-1-4666-9441-5.ch026

Wilson, J., & Roscoe, R. D. (2020). Automated writing evaluation and feedback: Multiple metrics of efficacy. *Journal of Educational Computing Research*, *58*(1), 87–125. https://doi.org/10.1177/0735633119830764

Woodworth, J., & Barkaoui, K. (2020). Perspectives on using automated writing evaluation systems to provide written corrective feedback in the ESL classroom. *TESL Canada Journal/Revue TESL du Canada*, *27*, 234–247. https://doi.org/10.18806/tesl.v37i2.1340

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing*, *27*(2), 147–170. https://doi.org/10.1177/0265532209349465

Xiao, Y., & Zhi, Y. (2023). An exploratory study of EFL learners' use of ChatGPT for language learning tasks: Experience and perceptions. *Languages*, *8*(3), 1–12. https://doi.org/10.3390/languages8030212

Yancey, K. P., LaFlair, G. T., Verardi, A. R., & Burstein, J. (2023). Rating short L2 essays on the CEFR scale with GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)* (pp. 576–584). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.bea-1.49

Yau, C., & Chan, K. (2023). University of Hong Kong temporarily bans students from using ChatGPT, other AI-based tools for coursework. *South China Morning Post*. https://www.scmp.com/news/hong-kong/education/article/3210650/university-hong-kong-temporarily-bans-students-using-chatgpt-other-ai-based-tools-coursework

Yoon, S., Miszoglad, E., & Pierce, L. R. (2023). *Evaluation of ChatGPT Feedback on ELL Writers' Coherence and Cohesion*. https://doi.org/10.48550/arXiv.2310.06505

Yun, J. (2023). Meta-analysis of inter-rater agreement and discrepancy between human and automated English essay scoring. *English Teaching*, *78*(3), 105–124. https://doi.org/10.15858/engtea.78.3.202309.105

Zaidi, A. H. (2016). *Neural Sequence Modelling for Automated Essay Scoring* [Unpublished Master's thesis, University of Cambridge]. https://www.cl.cam.ac.uk/

Zhai, X. (2022). *ChatGPT User Experience: Implications for Education*. https://doi.org/10.2139/ssrn.4312418

Zhang, M. (2013). Contrasting automated and human scoring of essays. *R & D Connections*, *21*(2), 1–11. http://www.ets.org/Media/Research/pdf/RD_Connections_21.pdf

Zhang, M., Yao, L., Haberman, S., & Dorans, N. (2020). Assessing scoring accuracy and assessment accuracy for spoken responses. In K. Zechner & K. Evanini (Eds.), *Automated Speaking Assessment: Using Technologies to Score Spontaneous Speech* (pp. 32–57). https://doi.org/10.4324/9781315165103-3

Zupanc, K., & Bosnić, Z. (2017). Automated essay evaluation with semantic analysis. *Knowledge-Based Systems*, *120*, 118–132. https://doi.org/10.1016/j.knosys.2017.01.006