Journal of Educational Technology & Online Learning

Volume 6 | Issue 4 | 2023

http://dergipark.org.tr/jetol

JETOL JOURNAL OF EDUCATIONAL TECHNOLOGY & ONLINE LEARNING

Assessing second-language academic writing: AI vs. Human raters

Vasfiye Geçkin [a] *, Ebru Kızıltaş [a], Çağatay Çınar [a]

[a] İzmir Democracy University, Türkiye

Suggested citation: 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.

Highlights

- This research draws attention to the level of agreement, strength, and direction of the relationship between academic writing scores assigned by a large language learning model (ChatGPT, in this paper) and human raters.
- This research highlights the finding that the level of agreement in the scores assigned by two of the human raters and ChatGPT is slight to fair and significant.
- This research emphasizes that the scores the five human raters assigned reveal a significant low-to-high positive correlation.
- This research suggests reliable scores could be obtained when ChatGPT scores are paired with those obtained from human raters to assess second-language college writing.

Article Info: Research Article

Keywords: *Second-language writing assessment, ChatGPT, higher education, generative AI, automated writing scoring*

Abstract

The quality of writing in a second language (L2) is one of the indicators of the level of proficiency for many college students to be eligible for departmental studies. Although certain software programs, such as Intelligent Essay Assessor or IntelliMetric, have been introduced to evaluate second-language writing quality, an overall assessment of writing proficiency is still largely achieved through trained human raters. The question that needs to be addressed today is whether generative artificial intelligence (AI) algorithms of large language models (LLMs) could facilitate and possibly replace human raters when it comes to the burdensome task of assessing student-written academic work. For this purpose, first-year college students (n=43) were given a paragraph writing task which was evaluated through the same writing criteria introduced to the generative pre-trained transformer, ChatGPT-3.5, and five human raters. The scores assigned by the five human raters revealed a statistically significant low to high positive correlation. A slight to fair but significant level of agreement was observed in the scores assigned by ChatGPT-3.5 and two of the human raters. The findings suggest that reliable results could be obtained when the scores of an application and multiple human raters are considered and that ChatGPT may potentially assist human raters in assessing L2 college writing.

1. Introduction

Assessment literacy refers to an understanding of fundamental principles underlying effective assessment practices (Popham, 2004). Starting with formulating an age and level-appropriate assessment tool, writing assessment literacy for many second-language (L2) writing instructors requires the process of editing, providing effective feedback, and evaluating the written work. The process remains difficult since it requires clear assessment criteria, relevant pedagogical knowledge, and overcoming time constraints and biases (White, 2009; Zhang, 2013). When human raters are involved, the writing scoring process inevitably becomes time-consuming, labor-intensive and inconsistent among and within the raters (Hua & Wind, 2019; Uto & Ueno, 2018). Despite the vast amount of time good teachers dedicate to ensure the validity and reliability of their writing assessment practices (Coombe, 2010), human-related factors, including

* Corresponding author: Department of Foreign Language Education, İzmir Democracy University, Faculty of Education, Türkiye.
e-mail address: vgeckin@gmail.com
This study was partly presented as a proceeding at the International Conference on Educational Technology and Online Learning held between 20-23 June 2023.

Doi: http://doi.org/10.31681/jetol.1336599
Received 2 Aug 2023; Revised 16 Oct 2023; Accepted 21 Oct 2023

fatigue, subjectivity, and inconsistency, can still interfere with the reliability of writing scores (Hussein et al., 2019; Peng et al., 2012). Understanding language assessment practices to achieve reliable and consistent scoring is crucial for teachers and students to prevent loss of time, money, motivation, and confidence (Crusan et al., 2016). In the face of all these challenges, the employment of Automated Essay Scoring (AES) tools can offer benefits such as reduced rating times and increased homogeneity in scoring student-written work in a range of educational settings. In spite of the benefits AES tools offer, they do not come free of shortcomings. For one, these tools did not prove to be effective, and students do not favor such assessment tools (Chen & Cheng, 2008). For another, these tools may overlook content development and coherence of the written work and may put students at an advantage when compared to scores assigned by human raters (Doewes, & Pechenizkiy, 2012).

Since the introduction of the first AES tool, *the Project Essay Grader (PEG)*, numerous (non)commercial AES applications have been developed as tools for writing assessment to assist writing instructors in the time-consuming task of writing assessment (Page, 1966; 2003). AES tools act as a measurement of technology and evaluate the written work. The major task of any AES system is to classify the essay types and turn each essay into a measurable unit in terms of style, word use, syntax structure, and content by using algorithms and statistical methods and assigning a numeric value. *PEG* starts with a training phase where the system is trained on a sample of 100-400 essays, and then it assigns scores to the essays based on estimated coefficients in the scoring stage. It exhibited a positive correlation of .87 with human raters (Dikli, 2006; Refaat et al., 2012). Non-commercial popular AES tools such as *the Tool for the Automatic Analysis of Cohesion (TAACO)* and *the L2 Syntactic Complexity Analyzer (L2SCA)* are used to explore the relationship between text cohesion and essay quality (Crossley & McNamara, 2016) and the syntactic and lexical complexity of the written work (Lu, 2010). Some well-known commercial AES tools such as *Intelligent Essay Assessor* (*IEA*; Pearson Education, 2010), *e-rater* (Educational Testing Service, n.d.), and *IntelliMetric* (Vantage Learning, n.d.) rely on assessing the writing style and the content of the essay. All these AES applications are reported to correlate with human rater scores to a certain extent (Gierl et al., 2014) even though they may fail to detect off-topic or plagiarized essays, which human raters have developed a critical eye for (Hoang, 2011). According to Lim et al. (2021) and Ifenthaler (2022) the most widely used AES tools are listed as PEG, IEA, e-rater and IntelliMetric.

One of the latest large language learning models (LLMs) based on artificial intelligence (AI) that teachers and students could benefit from is an AI form based on a generative pre-trained transformer (GPT) architecture named ChatGPT. The benefits and challenges ChatGPT offers have been well articulated in the field of higher education (Rasul et al., 2023) and foreign language teaching, learning, and assessment (Hong, 2023). On one hand, this form of AI could generate ideas, analyze and write tasks, and enhance the learning experiences of both the teachers and the students. It performs various language-related tasks, including generating texts, answering questions, doing translation, and producing responses that look like the human language, thanks to powerful algorithms structured within deep learning mechanisms (Lund et al., 2023). Moreover, the chatbot employs a text-based interface (Fraiwan & Khasawneh, 2023; Xames & Shefa, 2023). On the other hand, dissemination of false information, a lack of deep understanding, and potential academic integrity issues (Farrokhnia et al., 2023) cannot be overlooked, as anyone can sign in and start using the beta version of GPT-3.5 for free or the latest version GPT-4 on a monthly fee of $20 without any prior training (OpenAI, 2023).

This paper follows this latest fade in writing assessment and aims to contribute to the existing literature on the consistency of scores assigned by the AES systems and human raters in assessing the written work of college students. This case study examines the scores assigned to student paragraphs in an English as a Foreign Language (EFL) context and compares correlational relationships and levels of interrater agreement between ChatGPT-3.5 and five human evaluators. The paper is structured as follows. First, we will discuss previous work on using AI in writing assessment tools and introduce our study. After detailing the method and the results, we will offer a discussion of our findings with respect to the relevant work in the literature.

2. Literature

Using AI as an AES tool in assessing writing is not a new phenomenon. Some research findings report promising results in using algorithms in scoring student written work. Some AES products which utilize rubrics on a particular scoring algorithm are commercially available to score essays on the TOEFL or GMAT papers. In contrast, some other products have been developed and trialed in non-western countries. In the Malaysian context, for example, the *Intelligent Essay Grader (IEG)* has been used to score essays written as a part of the Malaysian University Test of English (Wong & Bong, 2021). A total of 459 English essays were analyzed via *IEG*, which introduced a rubric consisting of language appropriacy and accuracy and indicators of coherence, cohesion, language functions, and task fulfillment. The scores assigned to the essays were correlated with language and semantic analyses. Language features are cited as a better predictor of an essay's score than semantic features.

Similarly, the scores assigned to the essays (n=326) of Chinese undergraduate English majors were correlated with those scored by the application *WriteToLearn* and four human raters (Lui & Kunnan, 2016). The scoring application was more consistent and stringent than the trained human raters, but it still failed to score seven essays. In another study, a total of 3453 essays from American middle and high schoolers were evaluated by the AES system, *IEA*, and experienced human raters (n=19) through the given informative and persuasive prompts (Chan et al., 2023). It was found that the scores assigned by the application and the human raters were more consistent than the scores obtained from the human raters only. Lu (2019) examined the English writing level of Chinese students (n=114) through the automated writing evaluation system, *Juku*, which was reported to be less effective than human raters in giving a proper evaluation. Tsai (2012) analyzed the essays of Chinese senior high schoolers (n=953) and reported that the scores of human raters were more consistent than the conjoined scores of the human raters and the AES system. However, the system can assign consistent scores based on grammar errors, paragraphing, mechanics, and word count (Tsai, 2010).

Some research findings suggest that the scores assigned to written work could be more reliable when a human rater is matched with an application rather than when the scores of two trained human raters are considered. In one of the very early studies, Landauer et al. (2003) found that the scores assigned to the essays of middle schoolers by the *IEA* and human raters showed a correlation of .90. Hoang and Kunnan (2016) explored the effectiveness of the scoring program *MY Access!* with a comparison to human raters. The scores assigned to the essays, written by learners of English as a foreign language on three different prompts by *MY Access!*, were found to be moderately correlated with those of the human raters. Even though the application detected content words, it fell behind the human raters in recognizing these words within discourse. Similarly, Azmi et al. (2019) reported that the scores assigned to the essays (n=350) of Arab middle and high schoolers by the AES system, the *Automatic Arabic Essays Evaluators (AAEE)* met 90% accuracy with a correlation of more than .75 with the scores assigned by the human raters. The correlation between the human-application pairings was found to be higher than that of the human-human pairings. Mizumoto and Eguchi (2023) used ChatGPT to analyze 12.000 essays written by second-language learners of English on the TOEFL database. ChatGPT was introduced a writing rubric, on a scale from 0 to 10, which was developed in accordance with the linguistic correlates of human rating scores. The scores assigned by ChatGPT reflected three different levels of proficiency and could be used reliably with the scores of the human raters. Alikaniotis et al. (2016) analyzed the style and content of a dataset of nearly 13.000 essays written by middle and high schoolers. The observed correlation to the scores between the human raters and the augmented model of score-oriented word embeddings and long and short-term memory network was more than .90. Taghipour and Ng (2016) trained a neural network model to evaluate the style and content of the essays on the Kaggle dataset (https://www.kaggle.com/c/asap-aes7data) and correlated the scores obtained from their system with those from the AES system, *Enhanced AI Scoring Engine (EASE)*, which is publicly available. The correlation between the scores assigned by these two systems was .75, and the correlation between the scores assigned by the developed neural network and the human raters was .76. The similarity between the average scores of the human raters and the two AES systems seemed promising in developing a finely tuned scoring system.

Similarly, Dong and Zhang (2016) found that the correlation between the scores given by human evaluators and the neural network to the syntactic and semantic features of the essays was .75. Dasgupta et al. (2018) developed another neural network architecture that acted as the AES system. The model was fed on linguistic features such as part of speech, cohesion, lexical diversity, causality, and informativeness of the AES system to evaluate around 20.000 essays. The correlation between the AES tool and the human raters was more than .90. Overall, recent research suggests time and cost-effective results in using AI-based algorithms to score students' written work reliably and consistently without any frustration or exhaustion. Now we introduce our study.

3. Methodology

3.1. Research Design

The research design is correlational, in which five different scorers who were trained on the same rubric marked a single group of writing papers. Participant recruitment was made through convenience sampling among college freshmen students who were instructed in an English medium program. The paper explored the level of agreement and correlation between the grades assigned by human raters and the generative pre-trained transformer, ChatGPT-3.5. The specific research questions addressed were the following:

(i) What is the level of agreement between five human raters in assessing student paragraphs written in a second language?
(ii) What is the level of agreement between the human raters and ChatGPT in assessing student paragraphs in a second language?
(iii) What is the relationship between the scores assigned to student paragraphs in a second language by these human raters?
(iv) What is the relationship between the scores assigned to student paragraphs in a second language by the human raters and ChatGPT?

3.2. Data Collecting Tools

Three instruments were used to collect data: (i) two demographic questionnaires, one of which aimed at exploring students' perceived level of proficiency in second-language writing, and the other of which aimed at eliciting human rater information, (ii) a writing rubric to grade the papers and (iii) a paragraph writing task. The writing task comprised an age-appropriate and intriguing topic and some writing prompts (See Appendix A). A holistic rubric[1] was provided to five human raters[2] and ChatGPT-3.5 with explicit instructions to evaluate the given paragraphs based on the given criteria (See Appendix B). The rubric components were introduced to ChatGPT-3.5, and the human raters were given an hour of a standardization session to get familiarized with the rubric and the grading process.

3.3. Participants

The participants were Turkish advanced-level learners of English as a foreign language. They were all first-year college students who were qualified to pursue their departmental studies in a department where the medium of instruction was English. Nearly half of the participants had to complete a year of prep school to be eligible for departmental studies. After completing the required two-semester freshmen academic writing courses, the participants had prior experience writing in a second language at the sentence, paragraph, and

[1] Previous work suggests that the kind of rubric (e.g., holistic vs. analytic) used to rate written work, the number of raters who do the scoring and their level of experience affect the scores assigned to student written work. The reason why we used a holistic rubric was that all the human raters were experienced in assessing college level student writing and even when the raters were asked to use an analytic rubric, they may still tend to give overall grades based on their personal impression (Çetin, 2011:483).

[2] We would like to thank one of the anonymous reviewers who suggested that we needed to increase the number of the raters to five. As suggested in the literature, even though it is practically difficult, employing five raters could improve scoring reliability (Arslan Mancar & Gulleroglu, 2022:528).

text levels. The task was given at the end of the second semester of the 2023 Spring academic year. A total of 60 students were assigned the task, but 43 were evaluated by the GPT and the human raters. Of these participants, 20 were males, and 23 were females. Since all the classes were held online, the written paragraphs were submitted on a digital learning platform. The tables below summarize participant demographics:

Table 1a.

Student demographics

Characteristics	Mean (SD)	Range
Age	19.4 (1.6)	18-26
Age of onset in L2 writing	12.5 (3.7)	6-19
Age of perceived fluency in L2 writing	16.4 (2.1)	11-20

As seen in the table above, the mean age of the participants was 19, and the reported mean age when they started to write in English was before age 13. Their perceived mean age to have gained fluency in writing in English was reported to be age 16. Next, we introduce the rater demographics.

Table 1b.

Rater demographics

Characteristics	Gender	Age	Years of teaching experience
Rater A	male	28	7(7 years of teaching writing at college level)
Rater B	female	29	7(5 years of teaching writing at college level)
Rater C	female	36	14(13 years of teaching writing at college level)
Rater D	female	30	7(7 years of teaching writing at college level)
Rater E	female	34	11(6 years of teaching writing at college level)

As given in Table 1b, the raters came from a similar educational background and teaching experience. The recruitment criterion for the raters was to have a minimum of five years of teaching experience in college-level second-language writing.

3.4. Data Analysis

After receiving the grades assigned by the five human raters and ChatGPT-3.5, the scores were entered anonymously into Excel sheets. Then, tests of normality, correlation, and reliability were run on the statistical software Statistical Package for Social Sciences (IBM, 2017) to address the relevant research questions**.** Since the data did not meet the assumptions of normality and the sample size (n<50) was small, level of agreement (Kappa values) and correlation coefficients (Spearman's rho values) were reported to explore the relationship between the scores assigned by the chatbot and the human raters.

3.5. Validity and Reliability

The content validity of the instrument, the writing task, and the rubric was ensured through the opinions of three experts in the field. Necessary amendments were made to the rubric and the task based on the feedback received from the experts. The writing rubric was adapted from Rosmawan (2017). The instrument was piloted on five participants before the main data collection. Face validity was granted through the opinions of the focus group who participated in the pilot study. These five papers ensured that the human raters and the chatbot got familiarized with the writing rubric.

3.6. Research Procedures

Ethics clearance was obtained from the university board of ethics (ID: 2023/08-03). The participants were recruited through convenience sampling. First, the participants were given a demographic questionnaire. Then the students were asked to do a thorough reading on the topic and write a paragraph discussing the causes of unemployment among young people in developing countries. The students were given a week to do the task. A detailed writing rubric was used to score the students' written paragraphs. The rubric was introduced to ChatGPT-3.5, and the five human raters were trained to ensure the standardization of the writing scoring process. After the standardization session, which took around an hour, the human raters were given a week to score the student paragraphs according to the rubric that they were introduced. Research participation was voluntary, and the students were awarded extra points for their time and effort.

3.7. Findings

Recall that the study investigated the level of agreement, the strength and direction of the relationship between the scores assigned by the human raters and Chat GPT-3.5. The table below gives descriptive statistics of the scores assigned by different raters:

Table 2a.

Writing scores

Raters	Mean (SD)	Range
ChatGPT	76.98 (17.66)	40-100
Rater A	86.51 (13.61)	40-100
Rater B	73.49 (17.98)	10-100
Rater C	76.28(13.63)	50-100
Rater D	77.91 (13.55)	50-100
Rater E	64.30 (15.98)	15-90
Human (Avr)	75.70 (12.15)	10-100

As can be seen in Table 2a, the mean scores assigned to the student paragraphs by ChatGPT and the third human rater were similar. However, the first human rater gave higher scores to the students' written work. Since we had a small sample size (n=43), determining the distribution of the scores was important for choosing an appropriate statistical method. The results of a series of the Shapiro-Wilk test showed that the distribution of the scores assigned by ChatGPT (W= .885, *p*< .001), Human Rater A (W= .836, *p*< .001), Human Rater B (W= .915, *p*= .004), Human Rater C (W= .934, *p*= .018), Human Rater D (W= .918, *p*= .005), Human Rater E (W= .923, *p*= .008) and the Average Human Rater Scores (W= .944, *p*= .039), departed significantly from normality. Next, we report the level of agreement between the raters:

Table 2b.

Agreement across raters

Raters	Kappa Value(κ)	Approximate significance (*p*)	Level of agreement
ChatGPT vs. Rater A	.027	.653	Slight
ChatGPT vs. Rater B	.124	**.036***	Slight
ChatGPT vs. Rater C	.098	.090	Slight
ChatGPT vs. Rater D	.260	**.000*****	Fair
ChatGPT vs. Rater E	.101	.069	Slight
ChatGPT vs. Human (Avr)	.061	**.045***	Slight
Rater A vs. Rater B	-.052	.404	Slight
Rater A vs. Rater C	.155	**.018****	Slight
Rater A vs. Rater D	.023	.711	Slight
Rater A vs. Rater E	-.086	.062	Slight
Rater B vs. Rater C	.252	**.000*****	Fair
Rater B vs. Rater D	.263	**.000*****	Fair
Rater B vs. Rater E	.020	.762	Slight
Rater C vs. Rater D	.224	**.001*****	Fair
Rater C vs. Rater E	.069	.292	Slight
Rater D vs. Rater E	.019	.763	Slight

p<.05. p<.01, *** p<.001*

As shown in Table 2b, human raters showed some level of agreement among themselves and ChatGPT also exhibited some level of agreement with the human raters. More precisely, Human Raters A & C showed a statistically significant slight level of agreement. A significant and fair level of agreement was observed between Human Raters B &C, B&D and C&D. The scores assigned by ChatGPT showed a slight but statistically significant agreement with Human Rater B and the average scores assigned by the five human raters. The finding that the scores assigned by Human Rater D and ChatGPT showed statistically significant fair level of agreement is promising. Based on this outcome, we used a non-parametric test to determine the correlational relations between the scores given by different raters. Table 2c gives the correlation coefficients across raters:

Table 2c.

Correlation coefficients across raters

Raters	Spearman' rho (r_s)	Approximate significance (*p*)	Size of correlation
ChatGPT vs. Rater A	.138	.376	negligible
ChatGPT vs. Rater B	.049	.753	negligible
ChatGPT vs. Rater C	**.398****	.008	low positive
ChatGPT vs. Rater D	**.390****	.010	low positive
ChatGPT vs. Rater E	.128	.413	negligible
ChatGPT vs. Human (Avr)	.237	.130	negligible
Rater A vs. Rater B	**.362***	.017	low positive
Rater A vs. Rater C	**.471****	.001	low positive
Rater A vs. Rater D	**.370***	.015	low positive
Rater A vs. Rater E	**.619*****	.000	moderate positive
Rater B vs. Rater C	**.494****	.001	low positive
Rater B vs. Rater D	**.548****	.000	moderate positive
Rater B vs. Rater E	**.587****	.000	moderate positive
Rater C vs. Rater D	**.734****	.000	high positive
Rater C vs. Rater E	**.504****	.001	moderate positive
Rater D vs. Rater E	**.583****	.000	moderate positive

p<.05. p<.01, *** p<.001*

As summarized in Table 2c, Spearman's correlation analysis was run to determine the strength and direction of the relationship between the scores assigned by the raters. A significant low to moderate positive correlation existed between Human Raters A and B (r_s= .36, n = 43, *p*= .017), Human Raters A and C (r_s= .47, n = 43, *p*= .001), Human Raters A and D (r_s= .37, n = 43, *p*= .015), and Human Raters B and C (r_s= .49, n = 43, *p*= .001). A significant moderate to high positive correlation existed between Human Raters A and E (r_s= .62, n = 43, *p*< .001), Human Raters B and D (r_s= .55, n = 43, *p*< .001), Human Raters B and E (r_s= .59, n = 43, *p*< .001), Human Raters C and E (r_s= .50, n = 43, *p*= .001) and Human Raters D and E (r_s= .58, n = 43, *p*<.001). A high positive correlation existed between Human Raters C and D (r_s= .73, n = 43, *p*< .001). The correlational values between the individual human raters A, B and E and the Averaged Human scores and ChatGPT were positive but negligible. A significant low to moderate positive correlation existed between ChatGPT and Human Rater C (r_s= .40, n = 43, *p*= .008) and Human Rater D (r_s= .39, n = 43, *p*= .010).

4. Discussion

This study explored the correlational relationship and agreement between the scores assigned to student paragraphs by human raters and those by the generative pre-trained transformer ChatGPT-3.5. The first two research questions explored the degree of agreement between the human raters and the degree of agreement between the human raters and ChatGPT. The results of this study suggest that there was a slight to fair level of agreement in the scores assigned between four human raters. Similarly, there was a slight to fair level of agreement between the generative transformer and two individual human raters and the average scores assigned by the human raters. Two of the human raters showed a slight but statistically significant agreement with ChatGPT in the scores assigned to the student paragraphs. The next two research questions dealt with the correlational relationship among the human raters and between the chatbot and the human raters. All the human raters manifested a statistically significant positive correlation with each other. This study lends support to the findings, which suggest that for the sake of interrater reliability, a neural network algorithm can be matched with a human rater (Mizumoto & Eguchi, 2023) and that pairing the application and the human rater would yield more consistent results than pairing the human raters to evaluate student written work (Hoang & Kunnan, 2016). The scores assigned by ChatGPT showed a significant positive low to moderate correlation with the scores assigned by two of the human raters. The conjoined use of ChatGPT with human raters is in-line with the domineering view in the literature (Attali et al., 2013).

What is more, our study mirrored the findings of Tsai (2012) in the sense that the correlation between the human scorers was higher and more significant than the correlation between the system and the human raters. However, it needs to be pointed out that the AES system that Tsai (2012) used rated 13% of the essays unscorable, with comments such as "syntax problem" or "off-topic" (p. 332). In our study, ChatGPT rated all the paragraphs without any exceptions. Similarly, Wang and Brown (2008) found that the scores assigned to student essays (n=107) by the AES IntelliMetric and two human raters on a holistic rubric did not show a significant correlation; however, there existed a significant correlation between the scores assigned by the human raters and the AES tool at the sentence level. This may suggest that the AES tool could function more accurately than the human raters when assessing surface-level writing features at the sentence level.

On the other hand, human raters could give equal weight to writing features at the sentence level and other dimensions, including focus, development, organization, and mechanics. Thus far, neural network models extracted statistical-based, style (syntax)-based, and content-based features from the essays that go into AES analyses (Ramesh & Sanampudi, 2022). Amorim and Veloso (2017) identified two broad categories of features to be extracted by the AES systems. Domain features included simple linguistic features such as the number of pronouns and verbs, general features, on the other hand, encompassed grammar and style, organization and development, lexical complexity, and prompt-specific vocabulary use. Our rubric included mechanics, sentence level accuracy, development and organization, and vocabulary use. Using a more detailed rubric covering all the domains and general features with multiple dimensions could result in more consistency between the human raters and ChatGPT.

The study offers attractive findings for teachers and testers in utilizing a chatbot in addition to a human rater in assessing writing to receive immediate, reliable scores within a transparent and objectively scored rubric in a shorter period (Hussein et al., 2019). However, the deep learning models of GPT are still obscure since we are not that familiar with the processes that take place in the analyses (see Uto, 2021 for a review). Thus, the interpretability and explicability of these models need to be made available (Kumar & Boulanger, 2020) to the users, including teachers and test developers. For instance, human raters use *trins* which are variables of interest such as sentence structure, word choice and organization while grading student written work. AI-based tools, on the other hand, extract features termed as *proxes* which are variables that are chosen to proximate with the human trins (Bai et al., 2022). The issue arises when *proxes* cannot differentiate between surface level features such as spelling mistakes and sentence and essay length and deeper features that reflect the semantic and rhetoric dimensions of the written piece (Raković et al., 2021). More specifically, non-content-based features such as word and sentence counts are extracted at a higher rate than content-based features which focus on the semantic features of a text (Ramesh & Sanampudi, 2022). Finally, it needs to be noted that GPT-3.5 has been upgraded to GPT-4 by March 2023, which could result in improved efficacy and performance in writing assessment.

5. Conclusions and Suggestions

This paper suggests promising results with respect to matching a human rater with a chatbot, ChatGPT in this case, to obtain reliable and consistent results in assessing college-level writing. The study has limitations, including the limited number of participants, lack of different writing genres, lack of a more detailed writing rubric, and an absence of a second chatbot to average the obtained scores.

The findings of this study come with certain benefits of incorporating AI algorithms in second language writing assessment process. For instance, human raters are reported to overlook grammatical errors unless they are errors of pronoun use or word choice which impede understanding the intended meaning (Ma & Slater, 2015: 411). AES tools, on the other hand, offer a more consistent, expertise and bias-free evaluation of the student written work (Taghipour, 2017). Moreover, these tools save time in scoring essays and providing explanations in real-time (Kumar & Boulanger, 2020). They establish a fair, unbiased, and transparent scoring process which could decrease teacher burden in large-scale assessments (Rupp et al., 2019).

Although the findings of this study suggest that ChatGPT can be considered a promising and viable alternative that can act as an AES tool for writing assessment purposes, one needs to be cautious about the limitations and drawbacks of this AES tool before launching it for use by test developers and assessors. First, the model to be introduced to the chatbot needs to be finely tuned for improved accuracy (Sethi & Singh, 2022). Even though we did not employ such a model, there existed a certain level of correlation and agreement between the human raters and GPT. Introducing fine-grained measures of syntactic and lexical complexity and diversity would enhance the model and help attain reliable results in terms of writing quality. Second, the chatbot lacks the sense of a human rater (Deane, 2013), the shortcoming of which can be compensated by correlating the scores assigned by the AES tool and the human raters (Page, 2003). Third, measures such as the length of the written work could trick the AES tool, and a higher score could be assigned to a poorly written piece (Ifenthaler & Dikli, 2015). That is why the system cannot be trusted as the sole evaluator of written pieces since once the students figure out how it works, they can easily fool it (Dwivedi et al., 2023; Perelman, 2020). The last challenge is to program the AES tools and introduce rubrics that could measure human creativity. Further research needs to be conducted to explore the full capabilities and acknowledge the limitations of AI algorithms in order to assess the second language writing proficiency of larger numbers of students with a comparison of other AI-based scoring systems. Despite being criticized as a form of 'high-tech plagiarism' (EduKitchen, 2023), the technology is here to stay especially for Gen Z who suffers from nomophobia (Düzenli, 2021). What could be done further is to educate the teachers, students, and test developers on the ethical use of these AES tools.

References

Alikaniotis, D., Yannakoudakis, H., & Rei, M. (2016). Automatic text scoring using neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics: Volume 1 Long Papers* (pp. 715-725). Stroudsburg: Association for Computational Linguistics.

Amorim, E. & Veloso, A. (2017). A multi aspect analysis of automatic essay scoring for Brazilian Portuguese. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 94-102). Student Research Workshop: Association for Computational Linguistics.

Arslan Mancar, S., & Gulleroglu, H. D. (2022). Comparison of inter-rater reliability techniques in performance-based assessment. *International Journal of Assessment Tools in Education, 9*(2), 515-533.

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.

Azmi, A. M., Al-Jouie, M. F., & Hussain, M. (2019). AAEE–Automated evaluation of students' essays in Arabic language. *Information Processing & Management*, *56*(5), 1736-1752.

Bai, J. Y-H., Zawacki-Richter, O., Bozkurt, A., Lee, K., Fanguy, M., Sari, B. C., & Marin, V. I. (2022). Automated essay scoring (AES) systems: Opportunities and challenges for open and distance education. In *Proceedings of the Tenth Pan-Commonwealth Forum on Open Learning (PCF10)* (pp. 1-10). Canada Minutes of Congress.

Chan, K. K. Y., Bond, T., & Yan, Z. (2023). Application of an automated essay scoring engine to English writing assessment using Many-Facet Rush measurement. *Language Testing*, *40*(1), 61-85.

Chen, E. C-F., & Cheng, E. W-Y. (2008). Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes. *Language Learning and Technology*, *12*(2), 94-112.

Coombe, C. (2010). Assessing foreign/second language writing ability. *Education, Business and Society: Contemporary Middle Eastern Issues, 3*(3), 178-187.

Crossley, S. A., & McNamara, S. (2016). Adaptive educational Technologies for Literacy Instruction. New York: Routledge.

Crusan, D., Plakans, L., & Gebril, A. (2016). Writing assessment literacy: Surveying second language teachers' knowledge, beliefs, and practices. *Assessing Writing, 28*, 43-56.

Çetin, Y. (2011). Reliability of raters for writing assessment: Analytic-holistic, analytic-analytic, holistic-holistic. *Mustafa Kemal University Journal of Social Sciences Institute, 8*(16), 471-486.

Dasgupta, T., Naskar, A., Saha, R., & Dey, L. (2018). Augmenting textual qualitative features in deep convolution recurrent neural network for automatic essay scoring. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications* (pp. 93-102). Stroudsburg: Association for Computational Linguistics.

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing, 18*, 7-24.

Dikli S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning, and Assessment*, *5*(1), 1-36.

Dong, F., & Zhang, Y. (2016). Automatic features for essay scoring—an empirical study. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 1072-1077). Stroudsburg: Association for Computational Linguistics.

Doewes, A., & Pechenizkiy, M. (2012). On the limitations of human computer agreement in automated essay Scoring. In *Proceedings of the 14th International Conference on Educational Data Mining (EDM21)* (pp. 475-480). International Educational Data Mining Society.

Düzenli, H. (2021). A systematic review of educational suggestions on generation Z in the context of distance education. *Journal of Educational Technology & Online Learning, 4*(4), 896-912.

Dwivedi, Y.K., Kshetri, N., Hughes, L., ….Wright, R. (2023). Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, 1-63.

EduKitchen. (2023, January 21). *Chomsky on ChatGPT, education, Russia and the unvaccinated* [Video]. YouTube. https://www.youtube.com/watch?v = IgxzcOugvEI.

Educational Testing Service (n.d.). About the e-rater® scoring engine. Retrieved June 1, 2023, from https://www.ets.org/erater/about.

Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2023). A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education and Teaching International,* 1-15.

Fraiwan, M., & Khasawneh, N. (2023). A Review of ChatGPT Applications in Education, Marketing, Software Engineering, and Healthcare: Benefits, Drawbacks, and Research Directions. *arXiv preprint arXiv:2305.00237*.

Gierl, M., Latifi, S., Lai, H., Boulais, A., & Champlain, A. (2014). Automated essay scoring and the future of educational assessment in medical education. *Medical Education, 48*(10), 950-962.

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.

Hoang, G. T. L. (2011). Validating My Access as an automated writing instructional tool for English language learners (Unpublished master's thesis). California State University, Los Angeles.

Hoang, G. T. L., & Kunnan, A. J. (2016). Automated Essay Evaluation for English Language Learners: A Case Study of MY Access. *Language Assessment Quarterly*, *13*(4), 359-376.

Hua, C., & Wind, S. A. (2019). Exploring the psychometric properties of the mind-map scoring rubric. *Behaviormetrika*, *46*(1), 73-99.

Hussein, M. A., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: a literature review. *Peer Journal of Computer Science*, *5*, 208-224.

IBM Corp. Released 2017. IBM SPSS Statistics for Windows, Version 25.0. Armonk, NY: IBM Corp.

Ifenthaler, D. (2022). Automated essay grading systems. In O. Zawacki-Richter & I. Jung (Eds.), *Handbook of open, distance and digital education* (pp. 1–15). Springer.

Ifenthaler, D., & Dikli, S. (2015). Automated scoring of essays. In J. M. Spector (Ed.), *The SAGE encyclopedia of educational technology* (Vol. 1, pp. 64–68). Thousand Oaks, CA: Sage.

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

Lim, C-T., Bong, C-H., Wong, W-S., & Lee, N-K. (2021). A comprehensive review of automated essay scoring (AES) research and development. *Pertanika Science and Technology, 29*(3), 1875-1899.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474-496.

Lu, X. (2019). An empirical study on the artificial intelligence writing evaluation system in China CET. *Big Data*, *7*(2), 121-129.

Lui, S., & Kunnan, A. J. (2016). Investigating the application of automated writing evaluation to Chinese undergraduate English majors: A case study of WriteToLearn. *Computer Assisted Language Instruction Consortium*, *33*, 71-91.

Lund, B. D., Wang, T., Mannuru, N. R., Nie, B., Shimray, S., & Wang, Z. (2023). ChatGPT and a new academic reality: Artificial Intelligence-written research papers and the ethics of the large language models in scholarly publishing. *Journal of the Association for Information Science and Technology*, *74*(5), 570-581.

Kumar, V., & Boulanger, D. (2020). Explainable automated essay scoring: Deep learning really has pedagogical value. *Frontiers in Education*, *5*, 572367.

Ma, H., & Slater, T. (2015). Using the developmental path of cause to bridge the gap between AWE scores and writing teachers' evaluations. *Writing & Pedagogy*, *7*, 395-422.

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), 1-13.

OpenAI. (2023, March 14). GPT-4. Retrieved June 1, 2023, from https://openai.com/research/gpt-4

Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan*, *48*, 238-243. https://www.jstor.org/stable/20371545.

Page, E. B. (2003). Intelligent Essay Grade (PEG®) [Computer software]. https://www.measurementinc.com/products-services/automated-essay-scoring.

Pearson Education. (2010). Intelligent Essay Assessor (IEA)™ Fact Sheet [Fact sheet]. Retrieved June 1, 2023, from https://images.pearsonassessments.com/images/assets/kt/download/IEA-FactSheet-20100401.pdf.

Peng, X., Ke, D., Xu, B. (2012). Automated essay scoring based on finite state transducer: towards ASR transcription of oral English speech. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1* (pp. 50-59). Association for Computational Linguistics.

Perelman, L. (2020). The BABEL generator and E-rater: 21st century writing constructs and automated essay scoring (AES). *Journal of Writing Assessment*, *13*(1). https://escholarship.org/uc/item/263565cq

Popham, W. J. (2004). Why assessment illiteracy is professional suicide. *Educational Leadership*, *62*, 82-83.

Raković, M., Winne, P. H., Marzouk, Z., & Chang, D. (2021). Automatic identification of knowledge-transforming content in argument essays developed from multiple sources. *Journal of Computer Assisted Learning*, *37*, 903-924.

Ramesh, D. & Sanampudi, S. K. (2022). An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review*, *55*, 2495-2527.

Rasul, T., Nair, S., Kalendra, D., Robin, M., de Oliveira Santini, F., Ladeira, W. J., Sun, M., Day, I., Rather, R. A., & Heathcote, L. (2023). The role of ChatGPT in higher education: Benefits, challenges, and future research directions. *Journal of Applied Learning & Teaching*, *6*(1), 1-16.

Refaat, M. M., Ewees A. A., & Eisa, M. M. (2012). Automated assessment of students' Arabic free text answers. *International Journal of Intelligent Computing and Information Science, 12*(1), 213-222.

Rosmawan, H. (2017). The Analysis of students' writing before and after the implementation of ready-to-write approach. *Journal of Culture, Arts, Literature, and Linguistics*, *2*(1), 1-16.

Rupp, A. A., Casabianca, J. M., Krüger, M., Keller, S., & Köller, O. (2019). Automated essay scoring at scale: a case study in Switzerland and Germany. *ETS TOEFL Research Report Series*, 1-23.

Sethi, A., & Singh, K. (2022). Natural Language Processing based Automated Essay Scoring with Parameter-Efficient Transformer Approach. In *6th International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 749-756).

Taghipour, K. (2017). Robust Trait-Specific Essay Scoring using Neural Networks and Density Estimators. Unpublished Doctoral Dissertation, National University of Singapore, Singapore.

Taghipour K., & Ng, H. T. (2016). A neural approach to automated essay scoring. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 1882-1891). Stroudsburg: Association for Computational Linguistics.

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(*4), 328-335.

Tsai, M. (2010). Things that an automated essay scoring system can and cannot do. *In Proceedings of 2010 International Conference on ELT Technological Industry* (pp. 87-103). Pingtung, ROC: NPUST.

Uto, M. (2021). A review of deep-neural automated essay scoring models. *Behaviormetrika, 48*, 459-484.

Uto, M., & Ueno, M. (2018). Empirical comparison of item response theory models with rater's parameters. *Heliyon, Elsevier 4*(5), 1-32.

Vantage Learning (n.d.). Intellimetric®. Retrieved June 1, 2023, from https://intellimetric.com/direct

Wang, J., & Brown, M. S. (2008). Automated essay scoring versus human scoring: A correlational study. *Contemporary Issues in Technology and Teacher Education, 8(*4), 310-325.

White, E. (2009). Are you assessment literate? Some fundamental questions regarding effective classroom-based assessment. *OnCUE Journal, 3*(1), 3-25.

Wong, W. S., & Bong, C. H. (2021). Assessing Malaysian University English Test (MUET) Essay on Language and Semantic Features Using Intelligent Essay Grader (IEG). *Pertanika Journal of Science & Technology, 29(*2), 919-941.

Xames, M. D., & Shefa, J. (2023). ChatGPT for research and publication: Opportunities and challenges. *Journal of Applied Learning & Teaching, 6*(1), 1-6.

Zhang, M. (2013). Contrasting automated and human scoring of essays. *R & D Connections, 21*(2), 1-11.

Appendix A (Writing Task)

Instructions: Write a paragraph discussing the causes of unemployment among young people in developing countries. Write between 15-20 sentences. Make sure that you follow the guidelines of writing a paragraph. You may use the following points or any others that you wish to.

Possible causes:

- Rapid growth of population
- Existence of the defective (poor) education system
- Rural/ urban migration
- Use of inappropriate technology
- Wage policy problem
- Decline of expectations of recruitment
- Demographic issues

Appendix B (Rubric)

Instruction: Please evaluate the papers according to the rubric below. If you hesitate to decide on any exact score, please take the average score of two (adapted from Rosmawan, 2017).

Level	Criteria
Proficient (100 pts.)	- Writes the paragraph with clear topic sentence, fully developed ideas, and finishes with a concluding sentence. - Uses appropriate verb tense and a variety of grammatical and syntactical structures; uses complex sentences effectively; uses smooth transitions. - Uses varied, precise vocabulary. - Has occasional errors in mechanics (spelling, punctuation, and capitalization), which do not detract from meaning.
Fluent **(80 pts.)**	- Writes single or multiple paragraphs with main idea and supporting detail; presents ideas logically, though some parts may not be fully developed. - Uses appropriate verb tenses and a variety of grammatical and syntactical structures; errors in sentence structure do not detract from meaning; uses transitions. - Uses varied vocabulary appropriate for the purpose. - Has few errors in mechanics, which do not detract from meaning.
Expanding (60 pts.)	- Organizes ideas in logical or sequential order with some supporting detail; begins to write a paragraph. - Experiments with a variety of verb tenses but does not use them consistently; makes subject/verb agreement errors; uses some compound and complex sentences; has limited use of transitions. - Vocabulary is appropriate to purpose but is sometimes awkward. - Uses punctuation, capitalization, and mostly conventional spelling; errors sometimes interfere with meaning.
Developing (40 pts.)	- Writes sentences around an idea; some sequencing is present but may lack cohesion. - Writes in present tense and simple sentences; has difficulty with subject/verb agreement; run-on sentences are common; begins to use compound sentences. - Uses high frequency words; may have difficulty with word order; omits endings or words. - Uses some capitalization, punctuation, and spelling; errors often interfere with meaning.
Beginning (30 pts.)	- Begins to convey meaning through writing. - Writes predominantly phrases and simple sentences. - Uses limited and repetitious vocabulary. - Uses incorrect spelling.
Emerging (10pts.)	- Shows no evidence of idea development or organization. Uses simple words and expressions. - Copies from a model. - Shows little awareness of spelling, capitalization, and punctuation.