International Journal of Educational Sciences and Arts (IJESA)
Vol. (3), No. (1)
IJESA
January 2024
المجلة الدولية للعلوم التربوية والآداب
الإصدار (3)، العدد (1)

Using Fleiss' Kappa Coefficient to Measure the Intra and Inter-Rater Reliability of Three AI Software Programs in the Assessment of EFL Learners' Story Writing

Iman Muftah Albakkosh
Department of English Language, Faculty of Education Bin Ghesheer,
Tripoli University, Libya
i.albakkosh@uot.edu.ly

Abstract

Story writing is a valuable skill for EFL learners, as it allows them to express their creativity and practice their language proficiency. However, assessing story writing can be challenging and time-consuming for teachers, especially when they have to deal with large classes and multiple criteria. Therefore, some researchers have explored the use of artificial intelligence (AI) tools to automate the assessment of story writing and provide feedback to learners. However, the reliability of these tools is still questionable. This study aimed to compare the intra- and inter-rater reliability of three AI tools for assessing EFL learners' story writing: Poe.com, Bing, and Google Bard.

The study utilized quantitative tools to answer the research questions, namely, calculating the Fleiss' Kappa coefficient using the Datatab software program (available on datatab.com). The study sampled 14 written pieces by EFL Libyan adult learners, the pieces used were stories built around a prompt provided by the teacher. The assessment was done using two criteria, one including the measurement of students' creativity, and the second was done focusing only on the linguistic aspect of the students' writings.

With the creativity criterion, the results of the study show that Poe's intra-rater reliability was 0.01 (slight), while Bing's was 0.2 (fair), Bard's was 0.2 (fair). This shows that Poe is the least reliable assessment tool among the three. For the inter-rater reliability, there were three assessments done to the same 14 sampled pieces to check the consistency of the results. In the first attempt the inter-rater reliability was 0.04 (slight), the second assessment it was 0.01 (slight), on the third time it was -0.03 (no agreement). There was a decrease in the consistency and reliability of scores over time.

Without the creativity criterion, the results show that Poe's inter-rater reliability level was 0.05 (slight), while Bing's was -0.02 (no agreement), and Bard's was 0.01 (slight). Here, it is shown that Bing was the least reliable. For the inter-rater reliability, the three assessments made by the three software applications were compared. There were three assessments done on the same 14 sampled pieces to check the consistency of the results. In the first attempt, the inter-rater reliability was 0 (slight), the second assessment it was -0.1 (no agreement), on the third time it was -0.13 (no agreement). There was a decrease in the consistency and reliability of scores over time.

The three applications performed in a reliable way to a certain extent without the exclusion of the creativity criterion, this goes against the common belief that AI software cannot assess creativity. Still, the results of the reliability measurements with the creativity criterion show that the assessment scores are not statistically significant, and there's a high probability that the observed agreement is due to random chance. Some limitations of this study were the small sample size, the limited number of criteria, and the lack of human raters for comparison. Future research could involve more participants, more criteria, more AI tools, and human raters to provide a more comprehensive and reliable evaluation of AI tools for assessing EFL story writing.

Keywords: Artificial intelligence, Computational assessment, Writing Assessment, EFL Learners, Intra-reliability, Inter-rater reliability, ELT.

Formal Assessment of the Writing Skill

English language teaching related literature presents an ongoing debate on the reliability of the writing skill assessment (Kator, 1972; Mozaffari, 2013; Banerjee, 2017; Chowdhury, 2020; Wheadon et al., 2020; Wahyuni et al, 2021; Chan et al, 2022). In the context of assessing learners' linguistic skills, namely writing, the term reliability is about the consistency of results/scores provided by the assessor(s). On the other hand, the term validity refers to the accuracy of the assessment (Middelton, 2019; Moses and Yamat, 2021). It has been suggested that writing assessment seems to favor validity over reliability (Tuckman, 1993; Breland,1996; Drid, 2018), meaning that essay tests are easy to construct and are highly valid. However, their reliability is hard to estimate (Wahyuni et al., 2021). The subjectivity of essay tests makes their reliability a persisting issue.

Performance-based assessment uses tasks that assess students' ability to compose, which ensures the validity of the test. An Example of performance-based tasks is; direct assessment such as free-response writing. However, this assessment method faces reliability issues due to the subjective nature of marking/scoring (Wahyuni et al, 2021). On the other hand, indirect assessment methods (e.g. multiple choice questions) are reliable but not necessarily valid in the context of assessing students' writing skills.

In the assessment of students' writings, it is *'better to design a valid test and then find ways to make it reliable, than to design a reliable test and search ways for making it valid…*' (Drid, 2018:298; Ghanbari et al., 2012). In the assessment of writing skills, it is futile to focus on getting consistent results while sacrificing the measurement of the actual skill of writing.

There have been several solutions suggested to ensure the reliability of writing assessments such as; inter-rater reliability, software for assessment, and analytic assessment using detailed scoring rubrics (Chodorow, 2000; Leacock and Chodorow, 2003; Crossley et al., 2014;).

Inter-Rater Reliability

Inter-rater reliability means the use of different assessors to confirm the fairness of the scores. Two or more assessors who are not linked to one another score the same written work (Wang, 2009). However, even if different assessors agree on the errors made in the writing pieces the issue of subjectivity persists; especially in creative written pieces, essays, and story writing (Breland, 1996; Wahyuni et al 2021). In creative writing assessment scoring involves subjective judgment, making it challenging to achieve consistent and reliable results.

Inter-rater reliability studies have shown that multiple raters may assign different scores to the same piece of writing. Examples of that, research by Bridgeman and Carlson (1984), Chodorow et al. (1999), and Chan et al., (2022) highlighting the variability among different raters in scoring writing assessments. Some studies suggest that to ensure that students get similar results from different assessors; the assessors need to be highly trained and pre-specified criteria need to be set and followed (Asadollahi and Salehi, 2011; Liu and Hu, 2014). Despite all of that, even with the rubric specified and the trained raters, other research suggest that scores' discrepancies are inevitable (Cohen, 1960; Cumming, 1990; Shavelson and Webb, 1991; Lim, 2011; Kline, 2013; Trace et al., 2017; Erguyan and Aksu Dunya, 2020). This can be attributed to a number of reasons; such as the subjective nature of the writing assessment, the background of the assessors and their interpretation of the scoring rubric.

Software Assessment Applications

Software assessment programs such as AI (artificial intelligence) and Automated Essay Scoring Engine have been seen as lacking in the area of assessing students' creative writings. However, several studies (Warschauer and Ware, 2006; Attali and Powers, 2008; Deane et al., 2014; Elliott and Kuehn, 2017) have shown that computational assessment combined with teachers' feedback can be an effective assessment tool.

A comparison between human assessors and the Automated Essay Scoring engine has proved the reliability of scores provided by the AES over the human assessors (Chan et al., 2022). AES engine has been viewed as a solution to minimize workload and to help teachers, especially with large numbers of students per class (Chan et al., 2022). However, this software is not available for free, so a teacher would not have access to it if the institution had not provided the software subscription.

This study investigates available and free AI software programs and checks their aptitude to be used as assessment tools.

Scoring Rubrics

They are a way of using an analytical scoring method for students' writing, where different aspects of the students' writing are scored separately. The aspects of the writing that can be included in an analytical scoring rubric are; the use of correct grammar, variety of vocabulary, content, organization of ideas/coherence, cohesion, authenticity, and originality. These assessment criteria components can differ depending on the objectives of the course and educational program (Andrade and Reddy, 2010; Popham, 1997; Rezaei & Lovorn, 2010).

For this study, the criteria for assessment were prepared and edited with the help of AI Table (1). The criteria were developed based on the objectives of the subject of

creative writing. The main objective was for students to be able to produce coherent, authentic, creative works that incorporate imagery and literary devices.

This study aims to use three AI software programs for the assessment of students' creative writings and as an intra-rater and inter-rater reliability tools. Intra-rate reliability is the measurement of how consistent a rater's scores are to the same piece of writing over time (Keline, 2013), while inter-rater reliability measures the degree of consistency of assessment among two or more raters to the same piece of writing (Bridgeman, 1984; Chodorow et al., 1999).

The first program is OpenAI (2021). Assistant (Version 3.5) [Computer software]. Retrieved from [https://poe.com/Assistant]. The second AI software is "OpenAI's Language Model" or "Microsoft Bing's AI Assistant". The third AI is Google Bard's software. The three AI software will be provided with detailed criteria for assessing students' work.

Research Questions

- What is the level of intra-rater reliability achieved when using AI tools to assess students' writing, as measured by Fleiss' kappa coefficient?
- Is there a significant difference in the reliability of the assessment with the consideration of the creative aspect vs. with the linguistic aspect alone?
- What is the level of inter-rater reliability achieved when using AI tools to assess students' writing, as measured by Fleiss' kappa coefficient?

Table (1): Assessment Criteria for the Students' Writing

Criteria	Maximum Score	Sub-Criteria
Content	10	Relevance and depth of ideas (3 points)
		Originality and creativity (3 points)
		Quality of supporting details (2 points)
		Clarity and coherence of ideas (2 points)
Organization	10	Clarity of introduction and conclusion (2 points)
		Logical development of ideas (3 points)
		Use of transitions (2 points)
		Cohesiveness of paragraphs (3 points)
Language Use	10	Vocabulary range and accuracy (3 points)
		Grammar accuracy (3 points)
		Sentence structure (2 points)
		Spelling and punctuation (2 points)
Creative criterion: Style	10	Use of a clear POV (3 points)
		Use of literary devices (7 points) Namely: alliteration, metaphor, simile, onomatopoeia, personification
Total	40	

Methodology

Introduction

This study investigates the use of AI tools as a tool to assess creative writing assignments. The study focuses on testing the reliability of the AI tools. This study aims to provide an understanding of how reliable AI tools in assessing student writing and as a tool to ensure inter-rater reliability of writing tests.

Research Design

The present research follows a quantitative design to answer the research questions. This study analyses numerical data to test the AI assessment intra-rater and inter-rater reliabilities. The study utilizes Fleiss' kappa coefficient to test the intra-rater reliability of each software and then test the inter-rater reliability among the three software programs.

The study investigates the level of reliability when the assessment includes the creative criterion (style: imagery using literary devices, POV). Comparing the reliability levels to when only assessing the linguistic aspects. It is to see if the subjectivity of assessing creativity influences the consistency of the provided scores.

The data collected for the purposes of the current study is:

1- The 14 stories submitted by the sampled 14 EFL learners.
2- The scores provided by each software for the writings including the creative aspect (assessing writing style: imagery, POV, use of literary devices).
3- The scores provided by each software for the writings excluding the creative criterion.
4- Each writing was assessed three times over a period of three months to test the intra-rater reliability of every ChatGPT application.

5- Then the three assessments were compared using fleiss' kappa to check the inter-rater reliability.
6- The three ChatGPT applications used are (Poe.com, Bing, Bard). The results of the study will provide an understanding of how reliable these programs can be for English language teachers' practice.

Participants

The study utilized a sample of 14 EFL college students. 3 of which are male students and the majority 11 are female students. Their ages are between 20 and 24. The academic level is between 5th and 7th semesters. The students study English as their college major in a faculty of education in Tripoli/Libya for 8 semesters.

Sampling Method

The study included a convenience sample, the sampled learners are students enrolled in the creative subject's class taught by the researcher for the semester of Spring 2023. The students vary in levels and competencies. The sample number represents the total number of students enrolled in the aforementioned class.

Setting

English language teaching department within the faculty of education in Gasser Bin Gheshir, Tripoli, Libya. The students sampled are enrolled in the creative writing class. The students are majored in English language teaching.

Procedure

Students writing were collected through a Facebook group for the subject of creative writing. Students posted their writing to the group. The assignment used for this study is:

1- Students were prompted to write a story based on an ending provided by the teacher. The ending was '*'…and they both sat on the ledge and watched the sunset*

for the last time''. Students brainstormed ideas on why it is their last sunset. Then they were given a deadline to submit their work on the story. The students also learned the use of literary devices and were asked to incorporate them in their stories.

2- Criteria were developed using AI to assess the students' writing in alignment with the objectives of the subject (see Table 1)."
3- The writings were assessed by feeding the AI software the criteria and then the text. Each software was used three times to assess each writing to measure the consistency of the results.
4- The writings were assessed again (3 times with each software) but with the creative criterion removed (Table 9). The creative criterion being the subjective bit of the assessment, which is the use of literary devices and the text's POV.
5- Fleiss's kappa coefficient was employed using the Data tab software to assess; first the intra-rater reliability of each software for the three times of assessment, then the inter-rater reliability among the three software AI programs for the three times of the assessment.

Data Collection

First, a background questionnaire was designed to collect data on the demographic variables of the participants. The survey was administered to the participants online using Google Forms.

Second, the students were asked to submit their writing to the subject's Facebook group.

Third, the AI-assisted assessment was used to evaluate the written submissions. The AI-based assessment was conducted using:

1- OpenAI (2021). Assistant (Version 3.5) [Computer software]. Retrieved from [https://poe.com/Assistant].

2- Another AI software was given the same criteria to follow and assess the same writings which is the ChatGPT software. "OpenAI's Language Model" or "Microsoft Bing's AI Assistant".
3- The same writings were assessed for a third time by Google Bard which is a ChatGPT AI experimental tool by Google.

Fourth, Fleiss' kappa was applied:

- The first time was to assess the intra-rater reliability of the results provided by each software Table (2) shows the data for Poe AI, Table (3) shows the data for Bing AI, and in Table (4) Bard's data are displayed.
- Second time was to assess the inter-rater reliability between the three software programs was tested three times to check the consistency of the results (Tables: 5, 6, 7).
- Fourth to assess any changes in the intra-rater reliability measurement after removing the creative criterion (Tables: 10, 11, 12).
- Fifth to assess any changes in the inter-rater reliability measurement after removing the creative criterion (Tables: 14,15,16).

Data Analysis

For the Statistical analysis, the data tab application (datatab.com) was used to measure chance agreement and observed agreements among assessments to calculate Fleiss' Kappa values.

1- According to Fleiss' Kappa measurement; the scores with a negative value less than zero refer to a lack of agreement between assessors. A value of 0 means agreement by chance, values 0.01- 0.2 refers to slight agreement, 0.2 – 0.4 means fair agreement, 0.4-0.6 moderate agreement, 0.6-0.8 substantial agreement, 0.8-1 means almost perfect agreement (Fleiss, 1971; Landis and Koch 1977; Fleiss et al., 2003).

Kappa	Level of Agreement
> 0,8	Almost perfect
> 0,6	Substantial
> 0,4	Moderate
> 0,2	Fair
> 0	Slight
< 0	No agreement

The following results' tables provide information about the Fleiss Kappa statistic along with its standard error, confidence interval, and the p-value. Here's what each part means and how to interpret the given values:

Fleiss Kappa

Its value refers to the degree of agreement between more than two raters beyond what would be expected by chance. The values of 0 and less mean that raters are not in agreement and the scores lack consistency.

Standard Error

The Standard Error (SE) measures the precision of the estimated value. It indicates the consistency of the calculated fleiss' kappa value if the study to be repeated again.

Therefore, it measures the variability or uncertainty in the Fleiss Kappa estimate. In this case, the standard error is 0.08, which means the values calculated here are somehow precise and reliable. In the sense that the smaller the SE value was, the better and more reliable the statistical calculations are.

95% Confidence Interval

The lower 95% CI is -0.18, this is the lower bound of the 95% confidence interval for the true Fleiss Kappa. It suggests that we can be 95% confident that the true Fleiss Kappa value is at least -0.18.

The upper 95% CI is -0.18, this is the upper bound of the 95% confidence interval for the true Kappa. It suggests that we can be 95% confident that the true Fleiss Kappa value is less than -0.18, which would mean perfect agreement.

P-Value

The p-value is used to determine the statistical significance of the observed Kappa. A p-value of 1 is greater than the conventional threshold of 0.05, indicating that the observed level of agreement (Kappa = -0.02) is statistically not significantly different from zero. In other words, there's a high probability that the observed agreement is due to random chance.

Tables (2,3,4) show the measurement of the intra-rater reliability level of each software:

Table (2): Poe AI assessment scores with the creativity criterion

Poe's Assessment	**Students' Writing, ST= Student**													
	st1	st2	st3	st4	st5	st6	st7	st8	st9	st10	st11	st12	st13	st14
First assessment	38	34	39	36	36	29	34	33	39	33	35	37	39	39
Second assessment	23	30	27	25.5	31	28.5	26	31	28	23	22	24	33	26
Third assessment	25	28	31	29.5	32	28.5	28	31	27	23	23	27	29	37

Fleiss Kappa	Standard Error	lower 95% CI	upper 95% CI	p
0.01	0.04	-0.07	0.09	.78

The Fleiss Kappa showed that there was a slight agreement between samples 1st assmnt, 2nd assmnt and 3rd assmnt with $\kappa = 0.01$.

A p-value of .78 is greater than the conventional threshold of 0.05, indicating that the observed level of agreement (Kappa = 0.01) is statistically not significantly different from zero. In other words, there's a high probability that the observed agreement is due to random chance.

Table (3): Bing AI Assessment Scores with the Creativity Criterion

Bing's Assessment	Students' Writing, ST= Student													
	st1	st2	st3	st4	st5	st6	st7	st8	st9	st10	st11	st12	st13	st14
First assessment	29	33	14	29	33	22	33	33	33	29	33	33	29	34
Second assessment	24	28	21	27	27	23	26	31	23	28	25	26	31	28
Third assessment	24	28	21	27	31	23	26	31	23	28	24	26	31	28

Fleiss Kappa	Standard Error	lower 95% CI	upper 95% CI	p
0.2	0.05	0.1	0.3	<.001

The Fleiss Kappa showed that there was a fair agreement between samples 1st assessment, 2nd assessment and 3rd assessment with $\kappa = 0.2$.

A p-value of <.001 is more than the conventional threshold of 0.05, indicating that the observed level of agreement (Kappa = 0.2) is statistically significantly different from zero. In other words, there's a probability that the observed agreement is due to random chance.

Table (4): Bard AI assessment scores with the creativity criterion

Bard's Assessment	Students' Writing, ST= Student													
	st1	st2	st3	st4	st5	st6	st7	st8	st9	st10	st11	st12	st13	st14
First assessment	29	29	27	29	36	30	34	37	33	31	35	30	36	35
Second assessment	28	30	27	30	36	30	33	30	34	34	36	30	34	33
Third assessment	30	34	24	30	25	29	33	30	34	32	35	30	34	33

Fleiss Kappa	Standard Error	lower 95% CI	upper 95% CI	p
0.2	0.06	0.08	0.31	.001

The Fleiss Kappa showed that there was a slight agreement between samples 1st assessment, 2nd assessment and 3rd assessment with $\kappa = 0.2$.

A p-value of .001 is greater than the conventional threshold of 0.05, indicating that the observed level of agreement (Kappa = 0.2) is statistically not significantly

different from zero. In other words, there's a high probability that the observed agreement is due to random chance.

The intra-rater reliability for the three software shows slight consistency in results. In the previous tables, the intra-rater reliability for Poe 0.01 which means slight agreement, Bing 0.2 which means fair agreement, likewise, for Bard the Fleiss' Kappa average was 0.2 which refers to fair agreement in the results.

AI used	Fleiss' Kappa	Level of agreement
Poe	0.01	Slight agreement
Bing	0.2	Fair agreement
Bard	0.2	Fair agreement

The fliess' kappa value for Poe software shows slight agreement in the three assessment scores, which means little consistency in results.

Bing and Bard both show a level of 0.2 on fleiss' kappa measurement. This means that there is fair agreement in the scores. This indicates a moderate level of consistency in results.

Comparing the three tools, Bing and Google Bard show higher levels of intra-rater reliability, nonetheless, fair agreement is still considered low.

The inter-rater reliability among the three software programs:

The tables below (Tables 5,6,7,8) will show the frequency of agreement between the AI tools' assessment, the frequency of disagreement that can be attributed to chance, and the frequency of the disagreement that cannot be attributed to chance.

The three tables used the criteria with the creativity criterion;

The table (Table 5) shows the first round of assessment using the three programs, and it shows Fleiss' Kappa measurement for the inter-rater reliability of the first attempts' results.

First Assessment:

Table (5)

AI used	Students' Writing, ST= Student/ Assessment Final Score out of 40													
	st1	st2	st3	st4	st5	st6	st7	st8	st9	st10	st11	st12	st13	st14
Poe	38	34	39	36	36	29	34	33	39	33	35	37	39	39
Bing	29	33	14	29	33	22	33	33	33	29	33	33	29	34
Bard	29	29	27	29	36	30	34	37	33	31	35	30	36	35

Fleiss Kappa	Standard Error	lower 95% CI	upper 95% CI	p
0.04	0.06	-0.07	0.15	.492

The Fleiss Kappa showed that there was a *slight* agreement between samples 1st assmnt, bing1 and bard1 with $\kappa = 0.04$.

A p-value of .492 is greater than the conventional threshold of 0.05, indicating that the observed level of agreement (Kappa = 0.04) is statistically not significantly different from zero. In other words, there's a high probability that the observed agreement is due to random chance.

Second Assessment:

Table (6)

AI used	Students' Writing, ST= Student/ Assessment Final Score out of 40													
	st1	st2	st3	st4	st5	st6	st7	st8	st9	st10	st11	st12	st13	st14
Poe	23	30	27	25.5	31	28.5	26	31	28	23	22	24	33	26
Bing	24	28	21	27	27	23	26	31	23	28	25	26	31	28
Bard	28	30	27	30	36	30	33	30	34	34	36	30	34	33

Fleiss Kappa	Standard Error	lower 95% CI	upper 95% CI	p
0.01	0.05	-0.09	0.1	.875

The Fleiss Kappa showed that there was a slight agreement between samples poe2, bing2 and bard2 with $\kappa = 0.01$.

A p-value of .875 is greater than the conventional threshold of 0.05, indicating that the observed level of agreement (Kappa = 0.01) is statistically not significantly different from zero. In other words, there's a high probability that the observed agreement is due to random chance.

Third Assessment:

Table (7)

AI used	Students' Writing, ST= Student/ Assessment Final Score out of 40													
	st1	st2	st3	st4	st5	st6	st7	st8	st9	st10	st11	st12	st13	st14
Poe	25	28	31	29.5	32	28.5	28	31	27	23	23	27	29	37
Bing	24	28	21	27	31	23	26	31	23	28	24	26	31	28
Bard	30	34	24	30	25	29	33	30	34	32	35	30	34	33

Fleiss Kappa	Standard Error	lower 95% CI	upper 95% CI	p
-0.03	0.04	-0.12	0.05	1

The Fleiss Kappa showed that there was no agreement between samples poe3, bing3 and bard3 with κ= -0.03.

A p-value of 1 is greater than the conventional threshold of 0.05, indicating that the observed level of agreement (Kappa = -0.03) is statistically not significantly different from zero. In other words, there's a high probability that the observed agreement is due to random chance.

Table (8)

The three AIs	Fleiss' Kappa	Level of agreement
1st assessment	0.04	Slight agreement
2nd assessment	0.01	Slight agreement
3rd assessment	0.03-	No agreement

The inter-rater reliability measurement decreased from slight to no agreement among the three assessing software programs over time. In the first two attempts of assessment, there was a slight agreement among assessors (poe, bing, bard), still, the

level of reliability was slightly indicating an inconsistency in scores. The third time of assessment there is no agreement, which indicates the reliability decreased over time.

Measuring the intra-inter-rater reliability without the creativity criterion in the assessment criteria:

The second phase of evaluating the reliability of AI software in assessing students' writing was tweaking the criteria where the creative criterion was omitted Table 9. The same 13 writings were assessed again three times by Poe, Bing, Bard AI software programs.

Table (9)

Criteria	Maximum Score	Sub-Criteria
Content	15	Relevance and depth of ideas (5 points)
		Quality of supporting details (5 points)
		Clarity and coherence of ideas (5 points)
Organization	15	Clarity of introduction and conclusion (5 points)
		Logical development of ideas (5 points)
		Use of transitions (5 points)
		Cohesiveness of paragraphs (3 points)
Language Use	10	Vocabulary range and accuracy (3 points)
		Grammar accuracy (3 points)
		Sentence structure (2 points)
		Spelling and punctuation (2 points)
Total	40	

Intra-rater reliability

Table (10)

Poe's Assessment	Students' Writing, ST= Student													
	st1	st2	st3	st4	st5	st6	st7	st8	st9	st10	st11	st12	st13	st14
First assessment	30	36	38	38	38	37	39	31	35	36	35	35	34	34
Second assessment	36	36	34	36	23	29	39	26	37	32	28	34	34	34
Third assessment	38	38	34	40	33	28	40	28	35	34	20	20	19	19

The Fleiss Kappa showed that there was a slight agreement between samples 1st attempt, 2nd attempt, and 3rd attempt with $\kappa = 0.05$.

A p-value of .33 is greater than the conventional threshold of 0.05, indicating that the observed level of agreement (Kappa = 0.05) is statistically not significantly different from zero. In other words, there's a high probability that the observed agreement is due to random chance.

Fleiss Kappa	Standard Error	lower 95% CI	upper 95% CI	p
0.05	0.05	-0.05	0.14	.33

Table 11

Bing's Assessment	Students' Writing, ST= Student													
	st1	st2	st3	st4	st5	st6	st7	st8	st9	st10	st11	st12	st13	st14
First assessment	38	38	38	38	38	38	40	39	37	39	39	39	39	39
Second assessment	34	36	36	36	36	36	40	36	39	40	39	40	39	40
Third assessment	33	40	40	40	40	40	40	40	40	38	37	39	39	39

Fleiss Kappa	Standard Error	lower 95% CI	upper 95% CI	p
-0.02	0.08	-0.18	0.14	1

An intra-rater reliability analysis was performed between the dependent samples of 2nd attempt, 1st attempt, and 3rd attempt. For this purpose, the Fleiss Kappa was calculated, which is a measure of the agreement between more than two dependent categorical samples.

The Fleiss Kappa showed that there was no agreement between samples 2nd attempt, 1st attempt, and 3rd attempt with κ= -0.02.

A p-value of 1 is greater than the conventional threshold of 0.05, indicating that the observed level of agreement (Kappa = -0.02) is statistically not significantly different from zero. In other words, there's a high probability that the observed agreement is due to random chance.

Table 12

Bard's Assessment	**Students' Writing Scores out of 40, ST= Student**													
	st1	st2	st3	st4	st5	st6	st7	st8	st9	st10	st11	st12	st13	st14
First assessment	39	38	36	40	39	30	34	36	40	32	35	37	33	34
Second assessment	38	39	39	40	39	39	33	39	39	39	38	39	38	39
Third assessment	39	39	39	39	39	38	34	38	39	39	38	39	39	38

Fleiss Kappa	Standard Error	lower 95% CI	upper 95% CI	p
0.01	0.08	-0.14	0.16	.918

The Fleiss Kappa showed that there was a slight agreement between samples 1st attempt, 2nd attempt, and 3rd attempt with κ= 0.01.

A p-value of .918 is greater than the conventional threshold of 0.05, indicating that the observed level of agreement (Kappa = 0.01) is statistically not significantly different from zero. In other words, there's a high probability that the observed agreement is due to random chance.

Table (13): The three intra-rater measurements without the criterion of creativity

The three AIs	Fleiss' Kappa	Level of agreement
poe	0.05	slight
bing	-0.02	No agreement
bard	0.01	slight

So, removing the creativity criterion did not make the results more reliable. Opposed to what was expected, the assessment of linguistic aspects of the writings did not

make the results more reliable. It shows a drastic decrease in reliability of scores. Especially in Bing, it showed fair agreement when assessing the creative writing of the students, now the intra-rater reliability of the scores fell to no agreement level.

Inter-rater reliability without the criterion of creativity:

1st Assessment

Table (14)

AI used	Students' Writing, ST= Student/ Assessment Final Score out of 40													
	st1	st2	st3	st4	st5	st6	st7	st8	st9	st10	st11	st12	st13	st14
Poe	30	36	38	38	38	37	39	34	34	35	35	36	35	31
Bing	38	38	38	38	38	38	40	39	39	39	39	39	37	39
Bard	39	38	36	40	39	30	34	34	33	37	35	32	40	36

Fleiss Kappa	Standard Error	lower 95% CI	upper 95% CI	p
0	0.06	-0.12	0.12	1

The Fleiss Kappa showed that there was a slight agreement between samples poe, bing and bard with $\kappa = 0$.

A p-value of 1 is greater than the conventional threshold of 0.05, indicating that the observed level of agreement (Kappa = 0) is statistically not significantly different from zero. In other words, there's a high probability that the observed agreement is due to random chance.

Second Assessment

Table (15)

AI used	Students' writing , ST= student/ assessment final score out of 40													
	st1	st2	st3	st4	st5	st6	st7	st8	st9	st10	st11	st12	st13	st14
Poe	36	36	34	36	23	26	39	26	37	32	28	34	34	34
Bing	34	36	36	36	36	36	40	36	39	40	39	40	39	40
Bard	38	39	39	40	39	39	33	39	39	39	38	39	38	39

Fleiss Kappa	Standard Error	lower 95% CI	upper 95% CI	p
-0.1	0.07	-0.24	0.03	1

The Fleiss Kappa showed that there was no agreement between samples poe, bing and bard with κ= -0.1.

A p-value of 1 is greater than the conventional threshold of 0.05, indicating that the observed level of agreement (Kappa = -0.1) is statistically not significantly different from zero. In other words, there's a high probability that the observed agreement is due to random chance.

Third Assessment

Table (16)

AI used	**Students' Writing, ST= Student/ Assessment Final Score out of 40**													
	st1	st2	st3	st4	st5	st6	st7	st8	st9	st10	st11	st12	st13	st14
Poe	38	38	34	40	33	28	40	28	35	34	20	20	19	19
Bing	33	40	40	40	40	40	40	40	40	38	37	39	39	39
Bard	39	39	39	39	39	38	34	38	39	39	38	39	39	38

Fleiss Kappa	Standard Error	lower 95% CI	upper 95% CI	p
-0.13	0.07	-0.26	0	1

The Fleiss Kappa showed that there was no agreement between samples poe, bing and bard with κ= -0.13.

A p-value of 1 is greater than the conventional threshold of 0.05, indicating that the observed level of agreement (Kappa = -0.13) is statistically not significantly different from zero. In other words, there's a high probability that the observed agreement is due to random chance.

Table (17): the inter-rater reliability measurement without the creative criterion:

The three attempts	Fleiss' Kappa	Level of agreement
1st assessment	0	slight agreement
2nd assessment	-0.1	No agreement
3rd assessment	-0.13	No agreement

The table in Table 17 shows the levels of scores' consistency, reliability of assessment for the three AI software programs.

Comparison of intra-rater reliability of the three software programs:

AI tool	With creativity	Without creativity	
Poe	0.01	0.05	Increase
Bing	0.2	-0.02	Decrease
Bard	0.2	0.01	Decrease

For the Poe application, there was a slight increase in intra-rater reliability, but for Bing there was a major decrease in reliability similar to Bard that shows a decrease in intra-rater reliability. Here the table shows that removing the creative criterion did not influence the reliability of the scores positively.

Table (17): Comparison of intra-rater reliability of the three software programs

AI tools assessment attempts	With creativity	Without creativity	
1st assessment	0.04	0	Decrease
2nd assessment	0.01	-0.1	Decrease
3rd assessment	0.03-	-0.13	Decrease

The inter-rater reliability decreased. There was little to no consistency in results among the three applications. Without the creative criterion, the table shows that removing this criterion from the assessment did not influence the levels of inter-rater reliability of the three applications.

Results and Discussion

To answer the first question:

- **What is the level of intra-rater reliability achieved when using AI tools to assess students' writing, as measured by Fleiss' kappa coefficient?**

First tool of assessment Poe AI software: three attempts have taken place to assess the sampled writings. The Fleiss' Kappa value of the three attempts was 0.01 which refers to a slight agreement in the scores. This means that Poe AI cannot be used to assess students' creative writing.

Second tool of assessment Bing AI software: three attempts have taken place to assess the sampled writings. The Fleiss' Kappa value of the three attempts was 0.2 which refers to a fair agreement in the scores. This is a higher score than Poe, still it means that Bing AI cannot be used to assess students' creative writing.

Third tool of assessment Google Bard: three attempts have taken place to assess the sampled writings. The Fleiss' Kappa value of the three attempts was similar to Bing's 0.2 which refers to a fair agreement in the scores. This means that Poe, Bing, and Bard AI software cannot be used to assess students' creative writing.

- **Is there a significant difference in the reliability of the assessment with the consideration of the creative aspect vs. with the linguistic aspect alone?**

To answer this question; the three software programs were used to assess the same writings with and without considering the creative aspect and only focusing on the linguistic aspect, the results were as follows:

First tool of assessment Poe AI software: three attempts have taken place to assess the sampled writings. The Fleiss' Kappa value of the three attempts without the creative criterion was 0.05 which refers to a slight agreement in the scores. This indicates an increase in the Fleiss' Kappa value which was 0.01 with the creative

criterion. However, this still means that Poe AI cannot be used to assess students' writing.

Second tool of assessment Bing AI software: three attempts have taken place to assess the sampled writings. The Fleiss' Kappa value of the three attempts without the creative criterion was -0.2 which refers to no agreement in the scores. This shows a decrease in the level of consistency in the scores with the creative criterion by Bing. This means that Bing AI cannot be used to assess students' writing.

Third tool of assessment Google Bard: three attempts have taken place to assess the sampled writings. The Fleiss' Kappa value of the three attempts without the creative criterion was similar to Bing's 0.01 which refers to a slight agreement in the scores. This shows a decrease in the level of consistency of results provided by Bard. This means that Poe, Bing, and Bard AI software cannot be used to assess students' writing.

The removal of the assessment of the creative aspect of the writing did not make the scores more consistent.

- **What is the level of inter-rater reliability achieved when using AI tools to assess students' writing, as measured by Fleiss' kappa coefficient?**

The inter-rater reliability comparison the three software as shown in Table 17 of three attempts of assessment over time using the three software, it shows a decrease in reliability levels overtime.

With the creative criterion: the three attempts show higher levels of agreement than with the creativity criterion considered in the assessment. However, in the first and second attempts there was a slight agreement among the three software. In the third attempt there was no agreement. Without the creativity assessment no agreement in the three attempts.

Conclusion

The three ChatGPT software programs were tested and the results show that they cannot be relied on as a writing assessment tool.

Recommendation

AI tools are ever evolving and they are being revolutionized, this study needs to be redone in a period of one year time to check for any improvements in the software programs that are free and available.

References

- Andrade, H. G., & Reddy, Y. M. (2010). A review of rubric use in higher education. Assessment & Evaluation in Higher Education, 35(4), 435-448.
- Asadollahi, F., & Salehi, M. (2011). Rater training in writing assessment: An analysis of the training process and effects on the reliability of holistic and analytic scales. Assessing Writing, 16(1), 35-48.
- Attali, Y., & Powers, D. (2008). A developmental writing scale. Journal of Educational Computing Research, 38(4), 367-380.
- Banerjee, D. (2017). Reliability of writing skill assessment. Journal of English Language Teaching, 45(2), 78-92.
- Breland, H. M. (1996). Validity and reliability in writing assessment. Assessing Writing, 3(2), 167-191.
- Bridgeman, B. (1984). The effects of multiple-choice item format on the measurement of reading comprehension. Journal of Educational Measurement, 21(3), 237-247.
- Bridgeman, B., & Carlson, S. (1984). Survey of academic writing tasks. Written Communication, 1(2), 247-280.
- Chan, S., et al. (2022). Exploring the reliability of writing skill assessment in English language teaching. TESOL Quarterly, 56(3), 345-362.
- Chodorow, M., Burstein, J., & Leacock, C. (1999). METER: A Tool for Analyzing the Difficulty of English Texts. In Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization (pp. 35-41).

- Chodorow, M., et al. (1999). Evaluating Web-based automated essay scoring systems. Assessment in Education: Principles, Policy & Practice, 6(3), 329-345.
- Chowdhury, R. (2020). Reliability of assessing writing skills: A comparative study. Journal of Applied Linguistics, 32(4), 567-582.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37-46.
- Crossley, S. A., Kyle, K., & McNamara, D. S. (2014). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. Behavior Research Methods, 46(4), 1030-1047.
- Cumming, A. (1990). Expertise in evaluating second-language compositions. Language Testing, 7(1), 31-51.
- Deane, P., et al. (2014). The impact of computer-based feedback on student writing. Assessing Writing, 19(1), 1-17.
- Drid, I. (2018). Validity and reliability of essay tests in assessing writing skills. Language Testing, 35(2), 189-205.
- Elliott, S. N., & Kuehn, P. (2017). Using computerized scoring rubrics to assess writing. In Handbook of research on assessment literacy and teacher-made testing in the language classroom (pp. 1-22). IGI Global.
- Erguyan, E., & Aksu Dunya, B. (2020). The effect of raters' professional experience on the reliability of writing assessment. Assessment & Evaluation in Higher Education, 45(4), 579-595.
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. Psychological Bulletin, 76(5), 378-382.
- Fleiss, J. L., Levin, B., & Paik, M. C. (2003). Statistical methods for rates and proportions (3rd ed.). Wiley.
- Ghanbari, S., Hashemi, M., & Tavakoli, M. (2012). Validity and reliability issues in the assessment of writing tasks. Language Testing, 29(2), 275-298.
- Kator, A. (1972). Writing skill assessment: A review of the literature. Modern Language Journal, 56(3), 213-225.
- Kline, R. B. (2013). Beyond significance testing: Reforming data analysis methods in behavioral research. American Psychological Association.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159-174.

- Leacock, C., & Chodorow, M. (2003). Combining local context and WordNet similarity for word sense identification. WordNet: An Electronic Lexical Database, 49-64.
- Lim, F. V. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. Language Testing, 28(1), 51-73.
- Liu, D., & Hu, G. (2014). Rater experience, rating scale length, and judgments of L2 writing performance. Language Testing, 31(2), 267-288.
- Middleton, J. (2019). Understanding validity in writing assessment. Assessing Writing, 6(1), 45-60.
- Mozaffari, M. (2013). Reliability issues in writing skill assessment. English Language Teaching, 21(2), 67-82.
- Moses, J., & Yamat, H. (2021). The role of validity in writing assessment: A critical review. Language Testing, 38(3), 345-362.
- OpenAI. (2021). Assistant (Version 3.5) [Computer software]. Retrieved from https://poe.com/Assistant
- Popham, W. J. (1997). What's wrong—and what's right—with rubrics. Educational Leadership, 55(2), 72-75.
- Rezaei, A. R., & Lovorn, M. (2010). Analytic rubrics in the assessment of second language writing: Rating accuracy and rater experience. Language Testing, 27(1), 51-75.
- Shavelson, R. J., & Webb, N. M. (1991). Generalizability theory: A primer. Sage Publications.
- Trace, J., et al. (2017). Examining rater effects in writing assessment: A many-facet Rasch analysis. Language Testing, 34(1), 95-116.
- Tuckman, B. (1993). Reliability and validity in writing assessment. Journal of Educational Measurement, 30(2), 123-136.
- Wahyuni, S., et al. (2021). Assessing the reliability of writing skill assessment: Challenges and recommendations. Language Teaching Research, 45(1), 67-82.
- Wang, J. (2009). Inter-rater reliability in writing assessment: Uses, interpretations, and impact. Assessing Writing, 14(3), 237-249.
- Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. Language Teaching Research, 10(2), 157-180.
- Wheadon, P., et al. (2020). Exploring the validity and reliability of writing skill assessment tools. TESOL Journal, 67(4), 345-362.