English Language and Literature Studies; Vol. 13, No. 2; 2023
ISSN 1925-4768 E-ISSN 1925-4776
Published by Canadian Center of Science and Education

An Empirical Study on the Validity of the AES Systems Juku and iWrite for Continuation Writing Task Assessment

Ziqing Luo[1] & Si Luo[2]

[1] Jiangmen Polytechnic, Guangdong, China

Correspondence: Ziqing Luo, Jiangmen Polytechnic, Guangdong, China.

Received: March 2, 2023 Accepted: April 20, 2023 Online Published: April 25, 2023

doi:10.5539/ells.v13n2p46 URL: https://doi.org/10.5539/ells.v13n2p46

Abstract

The automatic English scoring (AES) systems are coming to the forefront of English learners' minds with their speed, accuracy and personalized feedback. However, fewer researchers have studied the validity of AES systems in assessing narrative texts such as continuation writing tasks. Therefore, this paper empirically investigates the scoring between two AES systems, Juku and iWrite, and the difference in scoring validity between these two systems and the teacher.

This study mainly uses a quantitative method. The subjects of the study were the continuation writing tasks scores of 30 senior high school students in a Chinese middle school. Each task was scored by a professional teacher, Juku and iWrite, all with a perfect score of 25. Then the scores were statistically analyzed using SPSS 26.0.

The results of the analysis showed that (1) iWrite was more consistent and correlated with manual scoring than Juku. (2) In terms of mean scores, the manual scores were significantly higher than the Juku and iWrite scores. (3) In terms of discrimination, the system scores were not as good as the manual scores, but the latter were more subjective. (4) In terms of accuracy and stability, the AES systems were higher than manual scoring. Therefore, learners can use the AES scoring system as a reference and practice narrative writing based on the system's feedback on grammar and the teacher's feedback on plot and content.

Keywords: validity, automatic scoring system, Juku, iWrite, narration, continuation writing task

1. Introduction

Effectively assessing students' English compositions is an important part of improving students' English writing skills. However, in the actual teaching in high schools, the timeliness of writing correction seriously lags behind due to limited human and material resources. In particular, the English college entrance examination paper has added the continuation writing task, which has greatly increased the weight of writing marks, thus further increasing the pressure of teachers to correct students' writing. In the face of this problem, the Internet has brought opportunities for systematic grading of English composition, and various automatic English scoring (AES) systems have emerged. Although the use of such systems can improve the speed of marking, unlike teachers who mark according to the syllabus, AES systems compare students' essays with a back-end corpus to form a score and a rubric. Therefore, the evaluation validity between different AES systems, and between AES systems and manual scoring are worth investigating.

The purpose of this study was to investigate the effectiveness of the AES system in evaluating narrative texts such as continuation writing tasks by comparing the ratings between the two AES systems, Juku and iWrite, and between the two systems and the teacher. Hopefully, this research can provide useful suggestions for students' learning of continuation writing, and also provide a useful reference for optimizing the AES system and promoting high school English writing teaching.

2. Literature Review

According to language test expert Li (2001), the scores of test results should have person separability, which means the distribution of scores should be spread out, and the main indicators include discrimination, standard deviation and difficulty. The most direct way to compare the consistency of automatic and manual scoring abroad is to collect and analyze three indicators: the maximum score difference, exact-plus-adjacent agreement and the Pearson correlation coefficient (Shermis & Burstein, 2003). This study focuses on synthesizing the two

theories to compare the scoring validity between AES systems, and between AES systems and manual ratings in terms of person separability, correlation and grade agreement.

The development and research of AES systems for essays abroad are more mature than those in China, and representative systems are PEG, IEA, and E-rater. Attali and Burstein (2006) found that E-Rater was not much different from manual scoring in evaluating the composition written by learners. Wang and Brown (2007) proposed that the mean of machine scoring is significantly higher than the mean of manual scoring. Attali, Lewis and Steier (2012) showed that as an automated evaluation system the scores given to essays were highly consistent with the manual scores, but the evaluation of essay content and higher levels of writing such as quality of meaning expression, development of ideas and organization was very limited.

In China, more than 1,700 universities and high schools use Juku and iWrite. Research on AES in China has increased and deepened in recent years, but empirical studies are limited. The research on narrative writing is even less. There are mainly three types of empirical studies on Juku (Yin, Jia, & Lin, 2017). The first type mainly investigates students' and teachers' satisfaction or specific opinions on the use of Juku by means of questionnaires or questionnaires combined with interviews. The second type is to investigate the effect of using Juku on the improvement of the subjects' English writing skills. The experimental group of the second-year students' writing performance was significantly better than that of the control group who did not use Juku after using it for one semester (Zhou, 2020). The last type is that the researchers used experimental data to verify the correlation and difference between the automatic and manual scores to demonstrate the validity and reliability of Juku scores, but the results were not very consistent. Gao (2021) studied the grading of argumentative essays of 104 first-year college students on Juku, and found that the automatic scoring results can effectively reflect the vocabulary complexity and fluency of students' compositions, and are significantly higher than manual scoring. This is different from the research result of Wang and Ye (2014), who believe that the fitting degree between Juku scoring and manual scoring is very low, only 0.45, and the score validity is far lower than the mainstream AES systems abroad.

In contrast, there are fewer empirical studies on iWrite. Only Li (2021) studied the evaluation validity of this system, and he only studied the task of "writing emails". His result showed that iWrite and human evaluation are not consistent in terms of content, language, and structure dimensions.

However, Wang and Chen (2019) argue that iWrite can provide students with a comprehensive intelligent review of their essays based on four dimensions: language, content, chapter structure, and technical specifications, and help students improve the relevance and coherence of their essays based on their grammatical knowledge. Although the number of studies on the Juku and iWrite has increased and deepened, few researchers have studied both of them. Only Wu (2020) briefly analyzed the similarities and differences between iWrite and Juku without any empirical data.

More importantly, few researchers have studied the scoring validity of the AES system for narrative texts. The theory of continuation writing was put forward by Wang (2012), a famous second language acquisition research expert. This theory combines foreign language input theory and output theory, and the continuation writing task under this theory has been incorporated into the English college entrance examination as a new question task, which is worthy of study. Therefore, the research questions are as follows.

1) Is there any consistency or correlation between Juku, iWrite, and manual rating in evaluating the continuation writing task of first-year students in high school?

2) Can the difficulty, discrimination, and standard deviation of the grading of Juku and iWrite reach the level of the manual rating in the evaluation of students' continuation writing task?

3. Research Methods

The test subjects of this study were 30 first-year students in a certain class of senior high school in Anhui Province. The students' continuation writings were selected from the county joint examination in the second semester. The test was held in March 2022, and after the test, the scores were graded by teachers in this school who had many years of experience in teaching writing in upper grades of English. After collecting the students' essays, the researcher entered them into Juku and the iWrite. In order to make the manual scoring and the systematic scoring comparable, the author converted the scoring criteria for the essays on Juku and iWrite into a continuation writing scoring format out of 25 points. After all the data was collected, it was statistically analyzed using SPSS 26.0 and EXCEL. The statistical methods included descriptive statistics, Kendall's harmony coefficient, Pearson's correlation analysis, and consistency analysis. Among them, Kendall's harmony coefficient was obtained by converting the scores of the three into order and then entered into SPSS 26.0 for analysis.

Prior to the study, consent was obtained from the class teacher and informed consent was signed. During data collection, confidentiality was maintained, unnecessary personal information was removed, students' names were replaced using codes, and data was stored in a secure cloud disk.

4. Research Results

The mean, minimum and maximum values of the 30 students' continuation writing scores were calculated using SPSS 26.0 and the results are shown in Table 1.

Table 1. Descriptive statistics

	N	Minimum	Maximum	Average
R	30	10.00	23.00	19.84
A1	30	12.00	19.50	16.58
A2	30	14.00	18.00	16.04

Note. R = manual rating, A1 = Juku rating, A2 = iWrite rating.

As shown in Table 1, the mean values of manual scoring for the continuation writing tasks are higher than those of the two AES systems, which indicates that the computerized scoring criteria used in Juku and iWrite are more stringent than the manual scoring criteria. It is worth noting that the mean scores of Juku and iWrite are similar, which indicates that the scoring criteria of them are similar.

As described in the review, person separability is mainly reflected in the difficulty, standard deviation, and discrimination metrics. The related metrics of system scoring and manual scoring are shown in Table 2.

Table 2. The person separability of the system rating and manual rating

	Difficulty	Standard Deviation	Discrimination
R	0.86	3.82	0.31
A1	0.85	1.87	0.16
A2	0.89	1.24	0.12

The results in Table 2 show that the difficulty value of manual rating (0.86) is slightly lower than that of AES rating (mean 0.87), but the difference is not significant. In terms of standard deviation, the manual rating (3.82) was significantly better than the system rating (1.56 on average), and the manual rating was more spread out, which means that there was a wider range of scores and the greatest individual differences. The rating of Juku was more dispersed than that of iWrite. The mean scores of the top 27% of students and the bottom 27% of students were used to calculate the discrimination. The AES (0.14 on average) was very poorly differentiated and lower than the manual rating (0.31). Juku's rating is slightly better than iWrite at distinguishing students' writing ability.

To derive the degree of consistency in rater ratings, the researcher calculated Kendall's Harmony Coefficient for the order. The results are shown in Table 3.

Table 3. Kendall Harmony coefficient

	Kendall's Wa	Asymp. Sig.
R-A1	0.675	0.098
R-A2	0.763	0.035
A1-A2	0.874	0.008

Lu and Huang (2010) believed that the consistency of the rank order of the ratings affects the credibility of the evaluation, and the degree of consistency of the ratings can be tested by applying Kendall's harmony coefficient to determine the credibility of the evaluation data and the validity of the evaluation activities. Generally, $p < 0.05$ and $w > 0.6$ is credible to prove that it is the consistency of the ratings. The results showed high but not significant enough agreement between manual rating and Juku rating. There was significant agreement between manual rating and iWrite rating and between Juku rating and iWrite rating, with higher agreement between AES systems than between the manual and the system.

To detect the correlation between AES systems rating and between the manual and the system rating, Pearson correlation tests were conducted and shown in Table 4.

Table 4. Person correlation

		R	A1	A2
R	Pearson correlation	1	.424*	.569**
	Sig. (Two-tailed)		.020	.001
	N	30	30	30
A1	Pearson correlation	.424*	1	.756**
	Sig. (Two-tailed)	.020		.000
	N	30	30	30
A2	Pearson correlation	.569**	.756**	1
	Sig. (Two-tailed)	.001	.000	
	N	30	30	30

*Note.**. At the 0.05 level (two-tailed), the correlation is significant; **. At the 0.01 level (two-tailed), the correlation is significant.

There was a significant correlation between them, but the lowest correlation was between the manual and Juku rating (0.424) and the highest correlation was between systems (0.756).

5. Discussion

The following observations were made in this paper.

Firstly, the manual rating is significantly higher than that of Juku and iWrite in terms of average score. This is similar to the results of Yin, Jia and Lin (2017) and different from Gao (2021) whose research uses argumentative essays. The controversy may be due to the fact that teachers pay more attention to the content and ideas of students' essays when they score, while the system can only pay attention to the difficulty of vocabulary use, sentence length, and sentence structure, and cannot "appreciate" essays like teachers, especially narrative essays, which pay more attention to plot development.

Secondly, on the index of person separability (Table 2), the difficulty judgments of the system rating was almost identical to the manual rating, which is consistent with the results of Huang (2016). However, in terms of standard deviation and discrimination, the system rating was significantly inferior to the manual rating. This indicates that the AES system is not as good as the person in terms of differentiating students' continuation writing abilities, but people are a bit more subjective. According to Huang (2016), a discrimination level of 0.3 or higher is ideal, while 0.25 or so is only an acceptable level. Thus, if the discrimination of manual scoring (0.31) is more desirable, the discrimination of systematic scoring (0.14 on average) needs to be improved urgently.

Thirdly, iWrite performed better than Juku and it was more reliable. In addition, in this study, the consistency between iWrite and Juku scoring was as high as 0.874, with a high stability. This indicates that the two systems have the same scoring criteria and are not affected by time, place, fatigue, mental state, etc., as people are.

Lastly, in terms of correlation, the correlation of the two systems was higher than the system rating and the manual rating, and met the requirement of Stemler (2004) that a correlation coefficient of about 0. 70 between raters is acceptable. Liang and Wen (2007) argued that according to this criterion, not only is the scoring validity of Juku unsatisfactory, but also the manual rating is no exception. What's more, Page, E. B. & Petersen N. S. (1995) reported that the average correlation between manual raters in foreign even in the 1995 experiment with the highest reliability was only 0.65. This shows that it is difficult to achieve a better score between manual raters, let alone between a system and a human. Therefore, under this criterion, iWrite rating is acceptable but Juku rating needs to be improved. This is not quite consistent with the results of Wang and Ye (2014), which proves that after 8 years, the corpus of Juku has been constantly updated and improved, and its rating parameters have also changed.

6. Conclusion

This study demonstrates to a certain extent the scientific validity of the AES systems of Juku and iWrite, which are more accurate and stable than manual scoring. Although the corpus of Juku is constantly updated, iWrite is more consistent and correlated with manual scoring than it. Both of them score the four aspects of vocabulary, sentence, chapter structure, and content relevance separately, which is not possible for manual scoring of high school continuation writing task, which is a positive implication for the application of both systems to English

writing teaching.

Thus, the AES scoring system serves as a writing aid that learners can use as a reference to write, practice, and revise more using its scores. However, it is not correct to rely only on the AES system to assess the quality of reading and subsequent writing. Teachers still need to provide multiple feedback on students' essays. Especially for narrative essays, teachers can focus their comments on plot and content when reviewing them.

The limitations of this study are that no qualitative research such as interviews were conducted and therefore the causal analysis for the data was not adequate. As well as the sample size was only 30, which was relatively small. In future research, firstly, the researcher can increase the sample size and combine qualitative and quantitative research to analyze the results more deeply. Secondly, there are also differences between the rating degrees of different types of essays, so future studies can use essays of different genres as corpus to verify the findings in this study.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater v.2. *Journal of Technology Learning & Assessment*, *4*(3), 3–30.

Attali, Y., Lewis, W., & Steier, M. (2012). Scoring with the computer: alternative procedures for improving the reliability of holistic essay scoring. *Language Testing*, *30*(1), 125–141. https://doi.org/10.1177/0265532212452396

Gao, J. (2021). Research on the scoring quality of English composition automatic scoring system of correction network. *Journal of Harbin University*, *7*, 102–105.

Huang, H. (2016). An empirical research into scoring validity of AES. *Journal of Zhejiang University of Technology* (Social Science), *1*, 89–93.

Li, Y. (2001). *The Science and Art of Language Testing*. Hunan Education Press.

Li, Y. (2021). *Research on Consistency between iWrite Automatic Scoring and Manual Scoring*. Unpublished master's thesis. Beijing Foreign Language University.

Liang, M., & Wen, Q. (2007). Comments and enlightenment on the automatic grading system of foreign compositions. *Foreign Language Teaching*, *5*, 7.

Lu, X., & Huang, Y. (2010). Scoring consistency test for grade sequence assessment in educational evaluation. *Jiangsu Education Research*, *13*, 47–48.

Page, E. B., & Petersen, N. S. (1995). The computer moves into essay grading: Updating the ancient test. *Phi Delta Kappen*, *76*, 561–565.

Shermis, M. D., & Burstein, J. C. (2003). *Automated essay scoring: a cross-disciplinary perspective*. Lawrence Erlbaum Associates Inc. https://doi.org/10.4324/9781410606860

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment Research & Evaluation*, *9*(4), 1–19.

Wang, C. (2012). Reading and writing—An effective way to improve the efficiency of foreign language learning. *Foreign Language World*, *5*, 6.

Wang, C., & Chen, R. (2019). Research on the teaching of follow-up writing based on the follow-up writing module of iWrite2.0 platform—Taking narrative text as an example. *Modern Communication*, *7*, 19–20.

Wang, H. (2016). An empirical study on the validity of the AES system for English writing. *Journal of Zhejiang University of Technology* (Social Science Edition), *1*, 89–93.

Wang, J., & Brown, M. S. (2007). Automated essay scoring versus human scoring: a comparative study. *Journal of Technology Learning & Assessment*, *6*(2), 29.

Wang, Z., & Ye, C. (2014). An empirical study on the online correction mode of English composition. *Journal of Changsha Railway University* (Social Science Edition), *1*, 161–163.

Wu, Y. (2020). A comparative study of intelligent reviewing systems for English writing—taking iWrite and Correction Network as examples. *Campus English*, *40*, 69–70.

Yin, X., Jia, Y., & Lin, Q. (2017). A comparative empirical study on the validity of automatic scoring of "Juku.com" and "Bingguo". *Journal of Hebei North University* (Social Science Edition), *1*, 91–96.

Zhou, X. (2020). *An empirical study on the influence of automatic composition scoring system on junior high*

school English writing. Unpublished master's thesis. Jiangxi Agricultural University.

Appendix

The scoring of teacher, Juku and iWrite

Student	Manual Rating	Juku Rating	iWrite Rating
1	19.5	65.5	62.8
2	19	51	58
3	19.25	62.5	60
4	21.5	68.5	70
5	21.5	66.5	66
6	21.5	64	65
7	21.5	68.5	65
8	21.5	61.5	60
9	21.5	78	68
10	21.5	76	72
11	22.5	69.5	67
12	22.5	68.5	68
13	22.5	65.5	66
14	22.5	74.5	70
15	22	75.5	72
16	22	61.5	58
17	22	68.5	71
18	22	77	62
19	22	67.5	66
20	23	70	69
21	23	76	68
22	23	57.5	56
23	23	65	68
24	14.5	66.5	59
25	10	65.5	60
26	12.5	61	58
27	14.5	69	64
28	16.5	70	62
29	11	51.5	58
30	16	48	56