International Journal of English Linguistics; Vol. 10, No. 6; 2020
ISSN 1923-869X E-ISSN 1923-8703
Published by Canadian Center of Science and Education

An Evaluation of China's Automated Scoring System *Bingo English*

Jianmin Gao[1], Xin Li[1], Peiqi Gu[1] & Ziqi Liu[1]

[1] Department of Linguistics, Zhejiang University, Hangzhou, China

Correspondence: Jianmin Gao, Department of Linguistics, Zhejiang University, Hangzhou, Zhejiang, 310 058, China. E-mail: jimmy_gao@zju.edu.cn

Received: June 30, 2020 Accepted: August 2, 2020 Online Published: August 6, 2020

doi:10.5539/ijel.v10n6p30 URL: https://doi.org/10.5539/ijel.v10n6p30

Abstract

The study evaluated the effectiveness of *Bingo English*, one of the representative automated essay scoring (AES) systems in China. 84 essays in an English test held in a Chinese university were collected as the research materials. All the essays were scored by both two trained and experienced human raters and *Bingo English*, and the linguistic features of them were also quantified in terms of complexity, accuracy, fluency (CAF), content quality, and organization. After examining the agreement between human scores and automated scores and the correlation of human and automated scores with the indicators of the essays' linguistic features, it was found that *Bingo English* scores could only reflect the essays' quality in a general way, and the use of it should be treated with caution.

Keywords: agreement, *Bingo English*, China's automated scoring systems, correlation

1. Introduction

With the development in education, the assessment of writing proficiency which is one of the essential components of communication has changed dramatically. In terms of the tasks involved in the assessment, the multiple-choice task was replaced by the constructed-response task which is more valid in examining students' language use (Qian, Zhao, & Cheng, 2020). The scoring way also witnessed some changes with the advances in computer science. Researchers expert in multiple areas such as second language acquisition (SLA), language assessment, and computational linguistics have been making efforts to develop an automated essay scoring (AES) system (Xi, 2010) to compensate the demerits of traditional human scoring which is costly, low-efficient and relatively low-consistent (Attali & Burstein, 2006; Dikli, 2006). Some AES systems can not only offer a holistic score but also present feedback on linguistic quality, in which case, the use of AES has gained great popularity in educational settings. Many large-scale high-stakes tests of English language have involved AES in scoring. For instance, *e-rater*, which is developed by Educational Testing Service (ETS) to score and evaluate student essays, has been employed for tests like Graduate Record Examination (GRE). To be specific, in the Analytical Writing section of GRE, each essay is scored both by at least one experienced human rater and *e-rater*. The final score will be the average of the scores from the human rater and *e-rater* if these two scores are highly consistent. If not, a second human score will be required, and the final score will be equal to the average of the human scores. AES is also used for complementing human scoring in the Test of English as a Foreign Language (TOEFL), in which essays are scored by human raters in terms of content quality and meaning as well as *e-rater* in terms of other linguistic features related to writing proficiency. Apart from the professional scoring applications like *e-rater* frequently used in high-stakes assessment, there are some other online applications that are designed to be publicly accessed and require no specific training to use (Lewis Sevcikova, 2018). These online scoring applications including *iWrite*, a commercial online program in China, are often used in low-stakes tests or just classroom tests in which the automated scores are directly reported as the assessment results. However, even though AES systems have many advantages in assessing student writing proficiency, different voices about their effectiveness and validity never stopped. In addition, since many AES systems lack transparency in terms of how they award the scores, which makes it even harder to investigate the validity issues, exploring the scoring criteria held by AES systems as a part of validation work is of vital importance.

2. Literature Review

In validating AES, some researchers presented crucial considerations with regard to the validity of AES scores, aiming to provide directions for future validation research. For example, Xi (2010) raised a series of validity

questions that were focused on potential construct underrepresentation and misrepresentation caused by the use of AES systems, the consistency between AES score features and theoretical expectation, the consistency among AES scores across different measurement contexts and other test or non-test indicators, and the impact of AES scores on test-preparation, language teaching and learning, and score-based decisions. Deane (2013) also discussed over the issue of construct underrepresentation by comparing the construct of writing skills with the construct measured by AES systems, concluding that AES systems merely focused on a limited construct which ignored the social and cognitive dimensions of writing skills. Moreover, Raczynski and Cohen (2018) shared some considerations on the calibration and test of AES systems.

Apart from putting forward theory-driven considerations for AES systems, some experts in language testing and assessment also conducted a systematical and critical review of the past studies. To be specific, Chapelle and Chung (2010) conducted an elaborate review of automated scoring. They first explained how AES systems for language assessment had become a unique research discipline and then appraised several major automated essay and speech scoring systems by examining the measured construct and human-machine scoring agreement. They also demonstrated the evaluation of an AES system within an interpretive argument and eventually pointed out that the research on validating AES systems and language processing could benefit each other in a significant way. Lewis Sevcikova (2018) also summarized the merits and demerits of both AES systems and human scoring based on a thorough literature review, suggesting that although online scoring applications could not replace human proofreaders, incorporating more technology into assessments was potential to improve the efficiency of instructors' teaching and testing and save them more chances to interact with students face to face. In addition, Yang, Buckendahl, Juszkiewicz and Bhola (2002) summarized three operational approaches in their review: (1) to demonstrate correlations or agreements between automated and human scores, (2) to investigate the relationship between automated scores and other indicators of students' language proficiency and (3) to explore the scoring process and mental models employed by AES systems. Their critical review not only confirmed the methodological advantages of previous studies but also shed light on future empirical studies.

In terms of the empirical research, many researchers validated AES systems using an argument-based approach. For instance, Chapelle, Cotos and Lee (2015) validated two AES systems frequently used in diagnostic assessment, *Criterion* and *Intelligent Academic Discourse Evaluator (IADE)*, demonstrating how the inferences identified in the interpretive argument framework outlined by Clauser, Kane and Swanson (2002) could be used to develop specific validity arguments for different AES systems. In addition, Enright and Quinlan (2010) justified an essay scoring process that used both human raters and *e-rater* within the validity argument developed for the TOFEL test by Chapelle, Enright, and Jamieson (2010). In the end, they stressed the importance of studying the relationship between human scores, automated scores, and other non-test indicators of writing proficiency, indicating that rater agreement was not the only source of evidence for validating human and automated scoring. Williamson, Xi and Breyer (2012) also discussed five inferences related to automated scoring with emphasis on different dimensions in validation research. Following their research, Weigle (2013) pointed out some considerations for the use of AES systems and then validated *e-rater* using an argument-based approach with the supportive data collected from another research (Weigle, 2011).

Some empirical research did not follow an argument-based approach, but they were still focused on different areas within a validation framework. For example, in collecting more criterion-related validity evidence, Weigle (2011) compared the *e-rater* scores of TOEFL ibt writing tasks with other non-test indicators like student self-assessment and non-test writing scores, finding that there always existed a moderate consistency. Ramineni, Trapani, Williamson, Davey and Bridgeman (2012a) also investigated the ability of both human and AES scores to predict the scores of GRE Verbal section and found that they had the equal power to predict. The similar results occurred in the studies of Attali (2007) and Ramineni et al. (2012b) as well, in which the human and automated scores were compared with scores of other sections in TOEFL. In addition to the studies of agreement between the automated score and other indicators of writing proficiency, there were also studies of correlation between the automated score and the estimated true score, which was the average of the scores given by a group of different raters to an essay. The common finding of such studies was that the correlation between automated and human scores was nearly the same as inter-rater correlation (Attali, 2007; Attali & Burstein, 2006; Cohen, Levi, & Ben-Simon, 2018). Apart from such analysis of correlation, researchers also tried to create a proper feature weights for the AES model, aiming to optimize the measurement properties and improve the reliability of AES scores (Attali, 2015; Bridgeman & Ramineni, 2017). Moreover, there were many other case studies except for the study of *e-rater*, such as *My Access* (Hoang & Kunnan, 2016), *WritetoLearn* (Liu & Kunnan, 2016), *iWrite* (Qian et al., 2020), *IntelliMetric* (Rudner, Garcia, & Welch, 2006), and the AES system used on the platform of massive open online courses (MOOCs) (Reilly, Stanford, Williams, & Corliss, 2014). Some

ingenious examinations of the validity of AES scores were also not rare. For instance, in the study of Powers, Burstein, Chodorow, Fowles, and Kukich (2002), writers of different backgrounds were encouraged to "trick" the *e-rater* by eliciting a score that was obviously higher or lower than the deserved one. The finding indicated that *e-rater*, and perhaps other AES systems, might fail to offer an accurate assessment result without the human proofreading.

Through on our literature review, some research gaps were detected. First, although many research found a high human-machine score agreement or positive correlation (Attali & Burstein, 2006; Dikli, 2006; Foltz, Streeter, Lochbaum, & Landauer, 2013; Rudner, Garcia, & Welch, 2006; Shermis, Burstein, Higgins, & Zechner, 2010; Shermis & Hamner, 2013), it was not the same in the study of Liu and Kunnan (2016), and human-machine score agreement should not be used as the sole validity criterion because it is an indicator of reliability rather than a perfect validation approach (Bennett & Zhang, 2016; Qian et al., 2020). Therefore, it is still necessary to investigate the human-machine score agreement and find other criterion-related validity evidence for AES systems. Second, even though many studies suggested that AES systems could not evaluate the higher-order aspects of writing proficiency because they could not read and understand like human raters (Attali, 2015; Attali & Burstein, 2006; Attali, Lewis, & Steier, 2012; Weigle, 2011; Zhang, 2013), there is a need for more empirical evidence. Third, most of the AES systems that researchers attended to were developed by institutions in America and unavailable in China, a country with a huge amount of EFL (English as a Foreign Language) learners. The need for higher efficiency in assessing students' writing skills has promoted a series of AES tools in China, among which *Bingo English* is one of the most popular systems in classroom writing assessment and the assessment of writing assignments. However, there are merely few studies concerned with the validation of China's flourishing AES systems, namely the studies of He (2013) in validating *Pigai* and Qian et al. (2020) in validating *iWrite*. Such studies are vitally important and indispensable in that they not only present educators the suggestions about the application of those AES systems, but also lay a foundation for the future improvement of those systems and the development of new applications. In order to bridge these research gaps, the current research tried to evaluate *Bingo English*, not just by relying on human-machine score agreement but also calculating the correlation between automated scores and the indices related to language learners' writing ability as defined by complexity, accuracy, and fluency (CAF), which have been widely used in quantifying writing proficiency of EFL learners. In addition, the research also paid attention to the role of the AES system in scoring content and organization to provide implications for language teaching, learning and assessment. Two research questions were addressed:

1) How is the agreement between scores produced by human raters and *Bingo English*?

2) How do scores produced by human raters and *Bingo English* correlate to the writing ability as defined by complexity, accuracy, fluency (CAF), content quality and organization?

About *Bingo English*

Developed by a team of American artificial intelligence experts together with the senior professors including Huilan Ying and Yongzhen Shao from the School of International Studies at Zhejiang University, *Bingo English* is a pioneering English composition grading system in China. The project started in 2004, and it was designed to help college students improve their English skills. Since it was developed, it has received a lot of attention and praise from the majority of teachers and students in universities. *Bingo English* has been on trial at Zhejiang University since 2007, and has gone through a small-scale trial to a large-scale operation with more than 4,000 people. According to the feedback of the teachers, the accuracy rate of its rating reached more than 95%, and it even exceeded the human rating in some aspects. Since then, the tool was gradually promoted to other high schools, colleges, and universities in Hangzhou city and even across the country. Today, there are nearly 1,000 colleges and universities in 39 provinces of China using *Bingo English*.

To ensure its high accuracy (up to 95%), the system is tested by leading English testing specialists every year. As displayed on the writing feedback webpage, *Bingo English* not only provides instant scores but also gives its users personalized feedback on vocabulary, grammar, style of writing, content and other aspects. Instant scores and feedback would be given as students revise their compositions, thus improving students' writing ability as well as independent learning ability. With improved readability, instructors can provide guidance from a macro level, which greatly improves instructor's work efficiency. In addition, the system would re-adjust the scores after the whole class summit their compositions so that instructors could have a clear picture of the ranking. Moreover, the administrator function enables instructors to create their own writing tasks, set word limits and deadlines, check students' progress, detect plagiarism and download their scores for further analysis.

3. Method

3.1 Materials and Participants

The research materials were 84 students' writings in an online English exam. The students were all non-English major freshmen and from two parallel classes of a college English course taught by the same instructor. The exam was held after two months of the start of the term to test what students had learnt during this period. For the writing section in the exam, there was an independent writing task with the topic of being a campus guide. Students finished the exam and handed in their paper online in different classrooms, using the computers provided by the college. The whole exam lasted for two hours. Students were supervised during the exam and not allowed to use any tools like the electronic dictionary. They were informed that their scores of the exam would take up at least 10% of the final grade of the course.

3.2 Data Collection

Two trained and experienced raters scored all the writings based on the 15-point holistic scoring scale of CET-4 (College English Test Band 4) writing section (See Appendix A), which is a national English test for college students, and the automated scores produced by Bingo English were also obtained. According to the scoring criteria of CET-4 writing section, there were five levels of students' writing quality including 2-point level, 5-point level, 8-point level, 11-point level, and 14-point level, with each score level containing a range of three scores (e.g., 2-point level contains a range from 1 point to 3 points). Before the independent formal scoring, the two raters had a discussion over the scoring criteria and then conducted a pilot scoring of 20 randomly selected essays. After reaching a high agreement ($r = .888$, $p < .01$), the raters independently scored the rest of the essays, and the average was the final score of each essay. In addition to human and automated scores, students' writing ability was also measured by linguistic features including CAF, content quality, and organization.

Complexity is composed of lexical and syntactic complexity. They were analysed by Coh-Metrix 3.0 (Graesser et al., 2004). The effectiveness of the indicators provided by Coh-Metrix in catching the textual characteristics has been confirmed in many published studies (McNamara et al., 2010). For syntactic complexity, we first selected all the indicators concerned with syntactic complexity according to the description of each indicator and then conducted a factor analysis using SPSS 17.0 to pick up the most related ones. Finally, three factors including seven indicators were obtained (See Table 1). For lexical complexity, the measure of textual lexical diversity (MTLD), which is "the mean length of word strings that maintain a criterion level of lexical variation" (McCarthy & Jarvis, 2010, p. 381) was used as the indicator because it is most insusceptible to the text length (McCarthy & Jarvis, 2010).

Table 1. Indicators of syntactic complexity

Factor	Abbreviation	Full name
Sentence level	MSL	Mean sentence length
	SS	Syntactic simplicity
	SSS (adjacent)	Syntactic structure similarity (adjacent)
	SSS (all)	Syntactic structure similarity (all)
Phrasal level	NPD	Noun phrase density
	VPD	Verb phrase density
Clausal level	LE	Left embeddedness

Accuracy was measured manually with the reference to the error types defined by Bardovi-Harling and Bofman (1989) and applied by Neumann (2014). Apart from the two primary types of error including the morphological error and the syntactic error defined in the previous study, the raters also identified other improper use of language such as the collocation error between an adjective and a noun (e.g., too much things). The human coders first had a pilot coding of 20 randomly selected essays, and after they reached a high agreement ($r = .927$, $p < .01$), they independently coded the rest of the essays. The raw incidence of errors may not be a proper indicator since it is quite possibly affected by the writing length, which purports that more words may contribute to more errors. In order to mitigate the influence of the writing length, we employed an error ratio (error number/total word number × 10) used in the study of Plakans et al. (2016).

Fluency was measured by the total number of words, as used in many other studies. As for content quality, with reference to the 'key points approach' adopted by Frost et al. (2011), the human raters first arrived at an agreement on the criteria of key points by analysing the writing instruction and reading through a random sample of essays. Since the students were required to introduce the campus environment in their essays, sentences used

to give instructions and building directions were counted as key points. Out of the same considerations as in measuring accuracy, a key-point ratio was adopted (key-point number/total sentence number). There was also a pilot coding of 20 essays, and after the human raters came to a high agreement (r = .941, p < .01), they started their formal coding independently. Through such an approach, both the degree of content richness and topic-relatedness could be measured.

Organization including coherence and cohesion was measured by the indicators provided by Coh-Metrix as well, and the most representative indicators were selected by a factor analysis using SPSS 17.0. Eventually, three factors containing 11 indicators from different dimensions were extracted (See Table 2).

Table 2. Indicators of organization

Factor	Abbreviation	Full name
Referential cohesion and coherence	NO (adjacent)	Noun overlap (adjacent)
	SO (adjacent)	Stem overlap (adjacent)
	NO (all)	Noun overlap (all)
	SO (all)	Stem overlap (all)
	LSA (adjacent)	Latent Semantic Analysis (adjacent)
Referential cohesion	AO (adjacent)	Argument overlap (adjacent)
	CWO (adjacent)	Content word overlap (adjacent)
	AO (all)	Argument overlap (all)
	CWO (all)	Content word overlap (all)
Causal relationship	DC	Deep cohesion
	CC	Causal connectives

3.3 Data Analysis

For the first research question, we first conducted an paired samples t-test to detect the difference between human and automated scores, and then an analysis of agreement using exact agreement percentage, adjacent agreement percentage, and Pearson's r to test the extent to which human and automated scores correlated (Attali & Burstein, 2006; Hoang & Kunnan, 2016; Qian et al., 2020). Scores were rounded for the calculation of exact and adjacent agreement percentage, since human raters only gave integral scores while *Bingo English* produced scores with one decimal place. For the second research question, analysis of correlation was done twice to detect how human and automated scores correlate with CAF, content quality, and organization, which are the linguistic features indicating writing proficiency. To be specific, the indicators of linguistic features were first correlated with human scores, and then automated scores. By comparing the correlation coefficients, we could go deeper to the nature of AES systems.

4. Results

For the first research question, Table 3 shows that human scores are higher than *Bingo English* scores in mean, minimum and maximum, and the result of the paired samples t-test suggests that such difference is statistically significant (p < .01, d = 1.299). This indicates that human raters tend to give higher scores than *Bingo English*. In addition, as Table 4 demonstrates, the exact agreement and adjacent agreement between human scores and Bingo English scores are quite low, with 13.10% and 34.52% respectively, and the correlation coefficient is .519, which is a sign of a moderate correlation.

Table 3. Descriptive statistics of human and *Bingo English* scores (N = 84)

	Mean	Standard Deviation	Minimum	Maximum
Human scores	10.64	1.17	5.00	12.50
Bingo English scores	9.14	1.14	4.40	12.20

Table 4. Agreement of human scores with *Bingo English* scores (N = 84)

Indicator	Value
Exact agreement (%)	13.10
Adjacent agreement (%)	34.52
Pearson's r	.519
Sig. (two-tailed)	.000

Note. Exact agreement: the proportion of times human scores exactly matched *Bingo English* scores; adjacent agreement: the proportion of times human scores were within one point of *Bingo English* scores.

Table 5 shows the results of the second research questions, and because neither human scores nor *Bingo English* scores were found to be correlated with the indicators of organization, these data are not presented in the table. According to the statistical outcomes, human scores are weakly correlated with the indicators of accuracy, fluency, and content quality, while *Bingo English* scores are weakly correlated with only one indicator of syntactic complexity (LE) and moderately correlated with the indicator of fluency. In addition, there is an unexpected result that human scores are negatively correlated with the error ratio, which is the indicator of accuracy, and there is also a trend approaching negative correlation between *Bingo English* scores and the error ratio (r = -.234).

Table 5. Correlation of human and *Bingo English* scores with the linguistic features (N = 84)

	Human scores		*Bingo English* scores	
	r	Sig (two-tailed)	*r*	Sig (two-tailed)
Complexity				
MTLD	.154	.161	.223	.042
MSL	-.027	.809	-.024	.825
SS	.227	.038	.147	.181
SSS (adjacent)	.064	.563	.056	.613
SSS (all)	.011	.924	.072	.516
NPD	.022	.843	-.104	.346
VPD	.008	.945	.053	.633
LE	.248	.023	.298*	.006
Accuracy				
Error ratio	-.358*	.001	-.234	.032
Fluency				
Word number	.436*	.000	.500*	.000
Content quality				
Key-point ratio	.417*	.000	.173	.115

Note. *| r | > .25 and p < .05.

5. Discussion

5.1 Agreement of Human Scores with Bingo English Scores

The outcomes of exact and adjacent agreement revealed that human scores and *Bingo English* scores were not always consistent. This corroborated with the study of *iWrite* (Qian et al., 2020), in which the exact agreement was 9% and the adjacent agreement was 34%, but conflicted with the studies of *My Access* (Hoang & Kunnan, 2016) in which the adjacent agreement amounted to 69.5% and *e-rater* (Enright & Quinlan, 2010; Powers, Escoffery, & Duchnowski, 2015), in which the exact agreement was 57% and 70% for each study, and the adjacent agreement was 98% and 99%, respectively. However, the correlation coefficient in the current study suggested that human scores were moderately correlated with *Bingo English* scores, which was consistent with the studies of *My Access* and *e-rater*, but inconsistent with the study of *iWrite*, which claimed that no correlation was found between human and *iWrite* scores (r = .037). In addition, both the current study of *Bingo English* and the study of *iWrite* found that human scores were higher than AES scores, and this was in contrast with the study of *My Access*, which indicated that automated scores were higher than human scores.

The comparison among different AES systems could possibly reveal not only the differences between AES systems in China and those primarily developed in America, but also the differences between AES systems used in low-stakes tests and those used in high-stakes tests. For one thing, the AES systems in China (*Bingo English* and *iWrite*) always used in a low-stakes context, might not produce the scores very close to human scores as the AES systems developed abroad and usually used in a high-stakes context did (*My Access* and *e-rater*), and practically, the scores produced by China's AES systems were lower than human raters. This was perhaps due to the different point systems. In this study and the study of *iWrite*, the essays were scored based on a 15-point scale, which is common in China, while the scale used in the studies of *My Access* and *e-rater* was either a 6-point or a 5-point scale. As was suggested by Ramineni and Williamson (2013), using a scale with few points included could lead to higher adjacent agreement, and this might even influence the result of Pearson's r according to the findings of Shermis and Hamner (2013). For another, as the AES systems used in America, *Bingo English* also produced scores that were correlated with human scores, but *iWrite*, which was another China's AES system did not show this feature. These findings suggested that *Bingo English* might be more

reliable than *iWrite*, and more efforts should be made to optimize AES systems in China to make it more reliable as the professional human scorers.

From the perspective of *Bingo English*, even though it did not produce the scores that shared a high exact or adjacent agreement with human scores, it was, by and large, able to judge the essay quality since the scores provided were correlated with the scores produced by professional raters. As for the significant difference between the two scores, one possible reason was that *Bingo English* scored essays based on a full-range scale with .1 point as an interval, unlike the human raters who scored with one point as an interval, and therefore it might score the essays in a stricter way. Another reason might lie in their different scoring procedures. With reference to a holistic scale, human raters tended to give reward scores based on an overall impression rather than deducting points in terms of the number of errors. In contrast, an essay was simply a bag of words for a machine, and they merely shared a stimuli-response relationship, which means the machine could only respond to the stimuli already built in the program. Therefore, with more attention paid to the global writing quality, the human raters might give higher scores than *Bingo English*.

5.2 Correlation of Human and Bingo English Scores with the Linguistic Features

As for complexity, only a weak correlation was found between *Bingo English* scores and one of the indicators of syntactic complexity, left embeddedness (LE), and no correlation was found between human scores and the indicators of linguistic complexity. One possible reason was that the complexity of essays did not show a significant difference among students. Since all participants were non-English major students in their first year of university, and the exam just happened two months after their entrance, their ability to use complex words and sentences might not differ very much. Another explanation was that linguistic complexity might not be the core of the scoring criteria both for human raters and *Bingo English*.

Unexpectedly, human scores were found to be negatively correlated with the error number, even though it was only a weak correlation. Also, *Bingo English* scores tended to have a negative correlation with the error number. On the one hand, this once again indicated that human scoring might be a top-down process, which means the essay was scored as a whole, so the details of an essay such as the incidences of errors might not contribute too much to the holistic score, instead, the number of error types might make a bigger difference. On the other hand, the result suggested that *Bingo English* was probably not very effective in recognizing linguistic errors, or linguistic accuracy was not a focus in its scoring criteria.

Both the human scores and *Bingo English* scores were correlated with the indicator of writing fluency, word counts. This result has also been found in many other studies. For example, Attali (2007) reported that both human scores and *e-rater* scores were moderately correlated with the word number with the correlation coefficient as .57 and .61, respectively. Also, Hoang and Kunnan (2016) suggested that the essay length was a strong predictor for both human scores and *My Access* scores by calculating a linear regression. In addition, in the study of Powers et al. (2015), a strong correlation (r = .87) was found between the word number and the scores provided by the certified human raters. All these results suggested the essay length was a correlate of writing quality which might be able to reflect fluent production and development, and most AES systems including *Bingo English* could capture this linguistic feature as a scoring criterion.

In terms of organization, neither human scores nor *Bingo English* scores were found to be correlated with the indicators of cohesion and coherence, and for content quality, it was only found to be weakly correlated with human scores. The results corroborated with the claim that AES systems were unable to evaluate the writing quality from the perspectives of content and organization because it could not truly understand the essay and evaluate the novelty or innovativeness of some ideas. However, in the current study, human scores were not correlated with organization, either, which conflicted the finding in the study of Powers et al. (2015) that there was a moderate correlation (r = .61) between the organization scores provided by *e-rater* and the scores given by the certified raters. The possible reason was that the introductory writing task employed in the current study might not so well discriminate students' organizational skills as the argumentative writing task did, which was used in the study of Powers et al., and therefore unlike the result in their study, the correlation between organization indicators and holistic scores was not shown in this study.

6. Conclusion

The findings of the study suggest that *Bingo English* as a teaching assistant tool can generally help instructors differentiate essays across different levels of quality, since the scores it provided in the study were basically correlated with the scores provided by the trained human raters. However, it is not proper to use *Bingo English* score as the only criterion to evaluate students' essays, for it was only correlated with the indicators of few linguistic features with higher-order aspects including content quality and organizational skills completely

excluded. Therefore, with the complementation of human scoring, instructors can use *Bingo English* to evaluate students' essay assignments or use it in some low-stakes tests, but employing it in high-stakes tests will be greatly risky. Also, the findings indicate that developers of AES systems in China should make constant efforts to refine the scoring model and make the systems more reliable and valid.

Several limitations should be acknowledged in this study. First, the writing genre of the essays collected in this study was single, thus the same results cannot be ensured when *Bingo English* is used to score other genres of writing such as narrative and argumentative writings. Therefore, future studies should collect essays of different genres as the research material and continue to validate the AES system. Second, the participants in this study were too homogeneous, and future studies ought to find people with a wider range in English proficiency as participants to test whether the same results would occur when *Bingo English* is used to score their essays. Third, the number of human raters was relatively small. Although both of the raters in this study were trained and experienced in scoring essays in a large-scale standardized test, their own understanding of the scoring criteria may still differ very much. According to Powers et al. (2015), there is an apparent difference in the correlation between the scores given by different human raters and the machine scores; therefore, in order to mitigate this issue, future studies should include a greater number of professional human raters to obtain more reliable human scores.

References

Attali, Y. (2007). *Construct validity of e-rater in scoring TOEFL essays* (ETS RR-07-21). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2007.tb02063.x

Attali, Y. (2015). Reliability-Based Feature Weighting for Automated Essay Scoring. *Applied Psychological Measurement*, *39*(4), 303–313. https://doi.org/10.1177/0146621614561630

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® v.2. *Journal of Technology, Learning, and Assessment*, *4*(3), 13–18. https://doi.org/10.1002/j.2333-8504.2004.tb01972.x

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

Bardovi-Harlig, K., & Bofman, T. (1989). Attainment of syntactic and morphological accuracy by advanced language learners. *Studies in Second Language Acquisition*, *11*, 17–34. https://doi.org/10.1017/S0272263100007816

Bennett, R., & Zhang, M. (2016). Validity and automated scoring. In F. Drasgow (Ed.), *Technology and testing: improving educational and psychological measurement* (pp. 142–173). New York: Routledge.

Bridgeman, B., & Ramineni, C. (2017). Design and evaluation of automated writing evaluation models: Relationships with writing in naturalistic settings. *Assessing Writing*, *34*, 62–71. https://doi.org/10.1016/j.asw.2017.10.001

Chapelle, C. A., & Chung, Y. (2010). The Promise of NLP and speech processing technologies in language assessment. *Language Testing*, *27*(3), 301–315. https://doi.org/10.1177/0265532210364405

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

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues & Practice*, *29*(1), 3–13. https://doi.org/10.1111/j.1745-3992.2009.00165.x

Clauser, B. E., Kane, M. T., & Swanson, D. B. (2002). Validity issues for performance-based tests scored with computer-automated scoring systems. *Applied Measurement in Education*, *15*, 413–432. https://doi.org/10.1207/S15324818AME1504_05

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

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, *18*(1), 7–24. https://doi.org/10.1016/j.asw.2012.10.002

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning, and Assessment*, *5*(1), 1–35.

Enright, M. K., & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater® scoring. *Language Testing*, *27*(3), 317–334. https://doi.org/10.1177/0265532210363144

Foltz, P. W., Streeter, L. A., Lochbaum, K. E., & Landauer, T. K. (2013). Implementation and applications of the intelligent essay assessor. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 68–88). New York, NY: Routledge.

Frost, K., Elder, C., & Wigglesworth, G. (2011). Investigating the validity of an integrated listening-speaking task: A discourse-based analysis of test takers' oral performances. *Language Testing*, *29*(3), 345–369. https://doi.org/10.1177/0265532211424479

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavioural Research Methods, Instruments, and Computers*, *36*, 193–202. https://doi.org/10.3758/BF03195564

He, X. (2013). Reliability and Validity of the Assessment by the Pigaiwang on College Students' Writings. *Modern Educational Technology*, *23*(5), 64–67. https://doi.org/10.18411/sciencepublic-04-11-2019-14

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. https://doi.org/10.1080/15434303.2016.1230121

Lewis, S. B. (2018). Human versus Automated Essay Scoring: A Critical Review Human versus Automated Essay Scoring: A Critical Review Lewis Sevcikova. *Arab World English Journal Arab World English Journal Arab World English Journal*, *9*(2), 157–174. https://doi.org/10.24093/awej/vol9no2.11

Liu, S., & Kunnan, A. J. (2016). Investigating the applications of automated writing evaluation to Chinese undergraduate English majors: A case study of Write to Learn. *CALICO Journal*, *33*, 71–91. https://doi.org/10.1558/cj.v33i1.26380

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*, 381–392. https://doi.org/10.3758/BRM.42.2.381

McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (2010). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes*, *47*, 292–330. https://doi.org/10.1080/01638530902959943

Neumann, H. (2014). Teacher assessment of grammatical ability in second language academic writing: A case study. *Journal of Second Language Writing*, *24*, 83–107. https://doi.org/10.1080/01638530902959943

Plakans, L., Gebril, A., & Bilki, Z. (2016). Shaping a score: Complexity, accuracy, and fluency in integrated writing performances. *Language Testing*, *36*(2), 1–19. https://doi.org/10.1177/0265532216669537

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2002). Stumping e-rater: Challenging the validity of automated essay scoring. *Computers in Human Behavior*, *18*(2), 103–134. https://doi.org/10.1016/S0747-5632(01)00052-8

Powers, D. E., Escoffery, D. S., & Duchnowski, M. P. (2015). Validating Automated Essay Scoring: A (Modest) Refinement of the "Gold Standard." *Applied Measurement in Education*, *28*(2), 130–142. https://doi.org/10.1080/08957347.2014.1002920

Qian, L., Zhao, Y., & Cheng, Y. (2020). Evaluating China's Automated Essay Scoring System iWrite. *Journal of Educational Computing Research*, *58*(4), 771–790. https://doi.org/10.1177/0735633119881472

Raczynski, K., & Cohen, A. (2018). Appraising the scoring performance of automated essay scoring systems—Some additional considerations: Which essays? Which human raters? Which scores? *Applied Measurement in Education*, *31*(3), 233–240. https://doi.org/10.1080/08957347.2018.1464449

Ramineni, C., Trapani, C., Williamson, D., Davey, T., & Bridgeman, B. (2012a). *Evaluation of e-rater® for the GRE® issue and argument prompts* (ETS RR-12-02). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2012.tb02284.x

Ramineni, C., Trapani, C., Williamson, D., Davey, T., & Bridgeman, B. (2012b). *Evaluation of e-rater® scoring engine for the TOEFL® independent and integrated prompts* (ETS RR-12-06). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2012.tb02288.x

Ramineni, C., & Williamson, D. M. (2013). Automated essay scoring: Psychometric guidelines and practices. *Assessing Writing*, *18*(1), 25–39. https://doi.org/10.1016/j.asw.2012.10.004

Reilly, E. D., Stafford, R. E., Williams, K. M., & Corliss, S. B. (2014). Evaluating the validity and applicability of automated essay scoring in two massive open online courses. *International Review of Research in Open and Distance Learning*, *15*(5), 83–98. https://doi.org/10.19173/irrodl.v15i5.1857

Rudner, L. M., Garcia, V., & Welch, C. (2006). An evaluation of IntelliMetric essay scoring system. *The Journal of Technology, Learning and Assessment*, *4*(4), 3–21.

Shermis, M. D., Burstein, J., Higgins, D., & Zechner, K. (2010). Automated essay scoring: Writing assessment and instruction. In E. Baker, B. McGaw & N. S. Petersen (Eds.), *International encyclopedia of education* (pp. 20–26). Oxford, England: Elsevier. https://doi.org/10.1016/B978-0-08-044894-7.00233-5

Shermis, M. D., & Hamner, B. (2013). Contrasting state-of-the-art automated scoring of essays. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: a cross-disciplinary perspective* (pp. 313–336). Mahwah, NJ: Lawrence Erlbaum Associates.

Weigle, S. C. (2011). *Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability* (ETS RR-11-24). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2011.tb02260.x

Weigle, S. C. (2013). English language learners and automated scoring of essays: Critical considerations. *Assessing Writing*, *18*(1), 85–99. https://doi.org/10.1016/j.asw.2012.10.006

Williamson, D., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, *31*(1), 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing*, *27*(3), 291–300. https://doi.org/10.1177/0265532210364643

Yang, Y., Buckendahl, C. W., Juszkiewicz, P. J., & Bhola, D. S. (2002). A review of strategies for validating computer automated scoring. *Applied Measurement in Education*, *15*, 391–412. https://doi.org/10.1207/S15324818AME1504_04

Zhang, M. (2013). Contrasting automated and human scoring of essays. *ETS R & D Connections*, *21*, 1–11. https://doi.org/10.1002/j.2333-8504.2013.tb02325.x

Appendix A

The Scoring Criteria of CET-4 Writing Section (Translated Version by the Researchers)

Level (points)	Description
13–15	Relevant; contents well developed without ambiguity; grammar, spelling, and punctuation almost free from errors.
10–12	Relevant; contents well organized; few errors in grammar, spelling, and punctuation, which do not interfere with comprehension.
7–9	Basically relevant; contents barely organized but not clear enough; frequent errors in grammar, spelling, and punctuation, some of which are serious errors that interfere with comprehension.
4–6	Basically relevant; contents not clear and badly organized; many serious errors in grammar, spelling and punctuation that interfere with comprehension.
1–3	Contents not clear and badly organized; fragmented sentences that abound with serious errors which interfere with comprehension.