International Journal of English Linguistics; Vol. 11, No. 2; 2021
ISSN 1923-869X E-ISSN 1923-8703
Published by Canadian Center of Science and Education

Unveiling the Scoring Validity of Two Chinese Automated Writing Evaluation Systems: A Quantitative Study

Jian Wang[1] & Lifang Bai[2]

[1] Department of English, Southwest Jiaotong University Hope College, Chengdu, China

[2] School of Foreign Languages, Hainan University, Haikou, China

Correspondence: Jian Wang, Department of English, Southwest Jiaotong University Hope College, Chengdu, Sichuan, 610 400, China. E-mail: 13076050176@163.com

Received: December 10, 2020 Accepted: January 10, 2021 Online Published: January 17, 2021

doi:10.5539/ijel.v11n2p68 URL: https://doi.org/10.5539/ijel.v11n2p68

Abstract

Computer Assisted Language Learning (CALL) has been a burgeoning industry in China, one case in point being the extensive employment of Automated Writing Evaluation (AWE) systems in college English writing instruction to reduce teachers' workload. Nonetheless, what warrants a special mention is that most teachers include automatic scores in the formative evaluation of relevant courses with scant attention to the scoring efficacy of these systems (Bai & Wang, 2018; Wang & Zhang, 2020). To have a clearer picture of the scoring validity of two commercially available Chinese AWE systems (*Pigai* and *iWrite*), the present study sampled 486 timed CET-4 (College English Test Band-4) essays produced by second-year non-English majors from 8 intact classes. Data comprising the maximum score difference, agreement rate, Pearson's correlation coefficient and Cohen's Kappa were collected to showcase human-machine and machine-machine congruence. Quantitative linguistic features of the sample essays, including accuracy, lexical and syntactic complexity, and discourse features, were also gleaned to investigate the differences (or similarities) in construct representation valued by both systems and human raters. Results show that (1) *Pigai* and *iWrite* largely agreed with each other but differed a lot from human raters in essay scoring; (2) high-human-score essays were prone to be assigned low machine scores; (3) machines relied heavily on the quantifiable features, which, however, had limited impacts on human raters.

Keywords: Chinese AWE systems, scoring validity, quantitative features, underlying causes, implications

1. Introduction

Writing proficiency constitutes a crucial component among EFL learning outcomes, but the evaluation task is notoriously taxing. Time and energy constraints make it impossible for even the most industrious and conscientious teachers to provide frequent writing assessments to a large writing class in the Chinese EFL teaching settings, which often leads to reduction in students' writing drills and their inadequate access to timely and detailed feedback (both quantitative and qualitative) which may very well facilitate learners' L2 development (Ziegler & Mackey, 2017). Moreover, as is indicated by Zhang (2013), human essay raters are subject to several errors and biases, such as severity/leniency, scale shrinkage, inconsistency, halo effect, stereotyping, perception difference and rater drift. These delicate issues have been partially eschewed by the application of Automated Writing Evaluation (AWE) systems which, including *Criterion*, *MY Access!* and WriteToLearn™ to name just a few, emerged in the wake of the Automated Essay Scoring (AES) engines like PEG™ (Project Essay Grader), IEA (Intelligent Essay Assessor), IntelliMetric, *e-rater* (Electronic Essay Rater). The extensive application of these systems into writing assessment has contributed to an increase in students' drill opportunities and in provision of timely scores and detailed feedback on content, organization, vocabulary and grammar (Dikli, 2006; Lee et al., 2009; Choi, 2014; Stevenson & Phakiti, 2014; Ranalli, 2018; Sarré et al., 2019). Therefore, AWE systems not only serve as scoring engines but also as Computer Assisted Language Learning (CALL) tools for users (Chen & Cheng, 2008; Grimes & Warschauer, 2008).

Research and development of Chinese AWE systems fares much later, compared with that of their foreign counterparts (especially the American ones) which date back to the 1960s. In the past decade, however, commercially available Chinese AWE systems such as *Bingo English*, *iWrite* and *Pigai* have been developed and adopted in EFL writing instruction across the country. These systems, as touted by their vendors, are characterized

by higher reliability and greater timeliness in providing scaffolding explanations and suggestions to activate learners' interlanguage knowledge. But some teachers' entire reliance on these systems to rate students' written products and the indiscriminate inclusion of the automated scores in the formative assessment may give rise to a fairness problem when there is still a cloudy picture of the effectiveness and authenticity of the machine scores (Bai & Wang, 2018; Wang & Zhang, 2020). The potential jeopardy of such practice may involve the elimination of the evaluative influence of teachers (Cheville, 2004) and the emergence of negative backwash effect. For instance, in order to get higher scores students tend to 'trick' the systems by intentionally catering to the assessment criteria of the machine, which may be irrelevant to the writing constructs (Powers et al., 2002).

To our knowledge, along with the wide integration of these systems into Chinese EFL writing instruction, only a handful of independent researchers and few developers or vendors in China have systematically released information about the AWE systems' scoring efficacy with respect to the comparability of these systems with human raters and the differences between human and machine scoring in essay evaluation. These questions are worth asking in view of the usefulness of machine scores. To bridge these gaps and to enrich the research field in Chinese AWE systems, this study takes a quantitative approach to address the scoring validity of *Pigai* and *iWrite*, two most successfully commercialized systems in China with a larger user base (with more than 600,000 users having subscribed *iWrite* and over 700 million essays rated by *Pigai*, as of December, 2020) and greater influence in English writing evaluation (with writing contests held by the vendors of both systems each semester). We take as a point of departure the extant literature in terms of the validity framework and the scoring validity of the AWE systems in China.

2. Literature Review

This section addresses the research framework that has been widely employed to validate the scoring performance of AWE systems, the research agenda with regard to Chinese AWE systems, the research gap and the endeavor of the present study.

2.1 Framework of AWE or AES Systems' Scoring Validity

Validity of measurement tools is part and parcel of language testing and psycho-metrics (Yang et al., 2002), which refers to the degree to which measuring tools or methods can accurately measure the measured things, or the accuracy and usefulness of these tools or methods (Zhang, 2017). Messick (1989) regarded the collection of abundant evidence as a guarantee for validity verification, and Weir (2005) incorporated scoring validity in the social cognitive framework for test validity. The enterprise for the validity of AWE or AES systems is mainly grounded in Kane's (2006) validity framework in which four dimensions are addressed: scoring, generalization, extrapolation and implication. By generalization, a relationship is set up between the observed machine scores and the scores to be expected from administering all possible similar essay tasks. In other words, it looks at the representativeness of the machine scores in comparison to scores from other possible essay tasks. Extrapolation validation examines the relationship between scores to be expected from administering all possible similar essay tasks and the scores from other measures in the domain of writing (i.e., various writing fields, such as academic writing, practical English writing, business English writing, and other fields related to the writing ability). Implication dimension tackles the relationship between the measure of writing ability from the assessment and subsequent interpretation for decision-making and prediction, so it plays a decisive role in language learning policies and strategies and a predictive role in language teaching practice, including syllabus formulation, teaching policy implementation and the like (Elliot & Williamson, 2013; Zhang, 2017).

The aforementioned three dimensions are important fields of inquiry in the validity study of AWE or AES systems. But the present study exclusively explores the scoring dimension, which addresses the relationship between the observed performance on the essay and the observed essay scores (the quality of the machine scores) (Elliot & Williamson, 2013). The enterprise for the scoring validity has been revolving around two directions: construct representation and score association. The former addresses the systems' effectiveness in measuring the constructs valued by human scorers, i.e., whether human raters and automated systems put a premium on identical or similar essay features (e.g., lexical and syntactic complexity, organization, etc.). This inquiry is thought to provide insights into how systems might be expected to approach human raters. For example, Deane (2013) reported that AES systems measured the essays' text organization, language and writing mechanics but provided inadequate evidence about the strength of argumentation or rhetorical effectiveness highly stressed in the scoring rubric for human scorers.

Score association concerns the consistency between automated and human scores. Exact agreement and exact-plus-adjacent agreement (EPAA) rates have gone mainstream as arguments for the scoring validity. For instance, the exact agreement rates of *e-rater*, IntelliMetric and *Criterion* ranged from 40% to 80% (Powers et al.,

2002; Vantage Learning, 2002; Shermis et al., 2008). The results of EPAA rates turn out to be more positive. Consider IntelliMetric. The figures in the studies generally stood at above 90% (e.g., Vantage Learning, 2002; Rudner et al., 2006) despite Powers et al.'s (2002) reporting on a rate of 65%. However, the use of agreement rate to indicate the correspondence between the machine and human scores has its limitations due to sensitivity to rating scales and the total number of research samples (Yang et al., 2002). To address this issue, diverse statistics (e.g., Pearson's correlation coefficient and Cohen's Kappa) are usually added to the statistical matrix. Take IntelliMetric for example anew: Rudner et al. (2006) got a correlation coefficient up to .80, while Wang and Brown (2008) found no correlation between AES and human scoring (r = .11, p > .05); Cohen's Kappa, which adjusts for chance agreement, differed at levels from .27 to .77 (e.g., Powers et al., 2002; Ramineni, 2013).

We based the present study on the scoring validity framework following Kane's (2006) representation of validity argument. In what follows, research relevant to the inquiry of Chinese AWE systems is reviewed.

2.2 Research on the Scoring Validity of Chinese English AWE Systems

With e-learning being highly commended in China (especially since the onset of the COVID-19 pandemic), AWE systems are having their heyday and are being more widely employed in Chinese EFL teaching settings. But the line of research on Chinese AWE systems remains a vast territory to be further exploited, whose effort, so far, has mainly revolved around the state of the art (Liang & Wen, 2007; Chen & Ge, 2008), the development of localized AWE systems (Li, 2009; Liang, 2011), and the effectiveness of applying these systems into writing instruction (Gu & Wang, 2012; Shi, 2012; Bai & Hu, 2017; Bai & Wang, 2019).

Compared to the American counterparts, the Chinese AWE systems are largely shrouded in mystery in terms of their scoring mechanism and efficacy, despite their extensive employment in writing assessment. So far scant attention has been given to the scoring validity of these systems, and only a handful of studies have dealt with this area, involving both construct representation and score association. The investigated systems include *Write On* (Wang, 2012), *Bingo English* (Gao et al., 2020), *Pigai* (He, 2013; Wang, 2016; Zhang, 2017; Bai & Wang, 2018; Xu, 2018), and *iWrite* (Li & Tian, 2018; Qian et al., 2020).

Wang (2012) investigated the scoring validity of *Write On*, an AWE system exclusively designed for the course *New Horizon College English*. This study sampled 200 essays and obtained a high human-machine correlation (r = .62) and a higher discrimination of the machine scores (i.e., the ability to distinguish high- and low-proficiency writers) than human scores. Besides, human raters and machine agreed more with each other in low-quality essays and the system tended to regard some off-topic essays with greater length as high-quality ones. Furthermore, the study indicated that the system focused more on content and language use and less on organization. But the conclusion was just drawn from the general comments made by the system and the linguistic features of the sample essays were not investigated. Gao et al. (2020) evaluated the scoring effectiveness of *Bingo English*, revealing low human-machine agreement (exact agreement rate = 13.10%, EPAA rate = 35.52%) and moderate correlation (Pearson's r = .519). This study also examined the correlation of human and machine scores with the indicators of the essays' linguistic features in terms of complexity, accuracy, fluency, content and organization, and found that machine scores could partially reflect the essays' quality. It must be pointed out, however, that the number of sample essays is too small (84 essays only) and that correlation analysis is not robust enough in corroborating the explanatory effect of the quantitative features on human and machine scores.

Most studies are related to *Pigai* but have produced mixed results. He (2013) obtained a higher human-*Pigai* correlation (r = .69) but found that the machine scores were significantly higher than human scores. This study further pointed out the ability of *Pigai* to diagnose some micro-structural errors (e.g., spelling and conventional grammatical errors) and its inability to evaluate macro-structural aspects stressed by human raters (e.g., the internal logic of the essay and relevance of the content). Wang (2016) inquired the scoring validity of *Pigai* from the perspectives of person separability, consistency and classification agreement (i.e., the percentage of essays whose machine-human score differences were within 3 points). The results showed that *Pigai* got more stable classification agreements (.86–.92) than human raters (.82–.96). Contrary to Wang (2012), Wang (2016) found a lower discrimination of machine scores, but it was concluded that the scoring validity of *Pigai* was so adequate as to satisfy the needs of English classroom writing tasks in spite of its relatively lower correlation coefficient (r = .53–.63) than the American systems. Zhang (2017) used the essays produced by 56 non-English majors as the research samples and found that *Pigai* highly agreed with two human raters in three rating tasks, with exact agreement rates ranging from 62.50% to 83.93%, EPAA rates from 98.21% to 100%, and correlation coefficients from .48 to .74. In contrast, Bai and Wang (2018) conducted a more detailed study, revealing *Pigai*'s fallibility to evaluate CET (College English Test) compositions due to its heavy reliance on the quantitative linguistic features.

But this study only analyzed a very small number of quantitative features and provided no criterion of feature selection. Xu (2018) sampled 70 CET-4 essays (College English Test Band-4, a high-stakes test in mainland China, which usually demands test-takers to finish writing an argumentative essay of no less than 120 words within half an hour), and indicated *Pigai*'s correct judgement of essay quality and partial representation of CET-4 writing constructs. However, the research design of this study is not without flaws. First, this study conducted a comparison between *Criterion* and *Pigai* scores to infer the scoring performance of the latter, disregarding the two systems' differences in scoring criteria and thus presumably compromising the comparability of both types of scores. Second, it is immensely untenable for this study to make inferences about the construct representation of *Pigai* just from its qualitative feedback. Third, the number of samples is actually thin for a solid conclusion to be drawn, feeding the suspicion of *Pigai*'s high inter-rater reliability and its ability to represent the CET-4 construct.

Two studies have investigated the scoring validity of *iWrite*, whose results are equally divergent. Li and Tian (2018) reported on high agreement and correlation between human scores and *iWrite* scores of 645 essays and concluded that *iWrite* was almost comparable to human raters (e.g., with the EPAA rate up to 97.98%). But this study was conducted by the developer and detailed information about the scoring performance of *iWrite* remains skeptical. Contrarily, Qian et al. (2020) showed unsatisfactory results of human-*iWrite* agreement, with the exact agreement between human and *iWrite* scores of total essays about 9%, EPAA rate 34%, Pearson's r .037 ($p > .05$), and Weighted Kappa -.02. The research design of this study is also problematic. First and foremost, both *iWrite* and human raters adopted a 15-point rating scale but their rating criteria were largely divergent and incomparable. Human raters employed an analytic scoring rubric based on the "ESL Composition Profile" (Jacobs et al., 1981), one that is different from *iWrite'*s scoring rubric. It is therefore natural to obtain low agreement between both types of ratings. Additionally, we believe it is problematic for this study to have adopted five fixed scores from each score level—2 points, 5 points, 8 points, 11 points and 14 points, namely, when two scorers assigned one essay 13 points and 15 points respectively both scores were then counted as 14 points. This practice might definitely inflate the inter-rater agreement.

2.3 Research Gap and the Endeavor of the Present Study

Taken together, extant research on the scoring validity of Chinese AWE systems has the following deficiencies. First, many studies are lack of comprehensiveness due to their exclusive emphasis on the association with human scores. Differences in the constructs valued by the machine and human raters have been given insufficient focus. Second, the studies centering on the construct representation drew general conclusions without statistical evidence and did not analyze the quantitative features of the texts in a deeper level. Third, some studies found a high machine-human agreement in scoring low-quality essays but failed to provide more evidence and explanations. Whether such a result can be applicable to both *Pigai* and *iWrite* demands further validation in this study. Fourth, studies touching upon horizontal comparison of different AWE systems are few and far between. Validity-related studies often involved only one system (*Pigai* mostly) and few examined two or more systems simultaneously, and so the results were incomparable most of the time, as the sample essays were from different writing populations. The machine-machine comparison is essential for finding the commonness or difference between AWE systems and for obtaining a clearer picture of their scoring performance. Although machine-machine comparison has not been widely explored in the literature, investigation into this aspect can shed light on the comparability of different AWE systems and contribute to identifying the common qualities or problems of these systems because the results from two systems with the same writing samples are more persuasive to showcase the machine-human difference. The horizontal data are equally helpful to infer the scoring mechanism of Chinese AWE systems due to limited information in this regard. So, we extend the scoring validity framework and also explore machine-machine differences or similarities in construct representation and score association. What should also be noted is that the corpus of most AWE systems in China (such as *Pigai* and *iWrite*) is constantly updated, and in response, its scoring validity might have changed accordingly. Therefore, more empirical studies are needed to follow up the changes.

In view of the shortcomings in the above-mentioned studies, the present study intends to unveil the scoring validity (construct representation and score association) of two Chinese AWE systems by extending the extant framework (i.e., involving both machine-human and machine-machine comparisons), calculating detailed statistics and gleaning more comprehensive quantitative linguistic features of each essay at the levels of accuracy, lexical and syntactic complexity, and discourse.

3. Methodology

3.1 Research Questions

Based on the review of previous work in the field, the following research questions were formulated:

(1) Are human, *Pigai* and *iWrite* scores congruent with one another?

(2) Are there essays inconsistently graded by human-*Pigai*, human-*iWrite* and *Pigai*-*iWrite* pairs? If any, what type do they belong to: low-, medium- or high-human-score essays?

(3) What are the machine-human and machine-machine differences or similarities in terms of construct representation?

3.2 Two Commercially Available Chinese AWE Systems: Pigai and iWrite

As mentioned previously, *Pigai* and *iWrite* are two most widely-adopted AWE systems customized for Chinese EFL learners. The former is a product of *Juku*, a search engine providing bilingual sentence examples. *Pigai* (meaning 'correction' in Chinese) bases its online service on cloud computing for automatically evaluating English compositions. It estimates the distance between the submitted compositions and the learner corpora, and generates essay scores and automatic feedback simultaneously.

iWrite is jointly developed by Foreign Language Teaching and Research Press of China and National Research Center of Foreign Language Education, Beijing Foreign Studies University. It is devised on the basis of in-depth research on L2 writing, corpus, natural language processing, machine learning, etc. As its vendor claims, *iWrite* evaluates essays in four dimensions: language, content, text structure and mechanics (spelling, capitalization, punctuation, etc.), and also pays attention to in-depth teacher-student interaction in the teaching and learning process.

3.3 Participants and Materials

Four hundred and eighty-six second-year non-English majors from a certain university in Southwest China participated in this study. The participants majored in *Civil Engineering*, *Accounting*, *Marketing* and *Human Resources* and took the compulsory course *College English IV*. This university offers no special English writing programs to students and the English writing skills and strategies are integrated into College English courses which last for four semesters. Due to time and energy constraints, English teachers often require students to write on *Pigai* platform.

In the present study, all the participants produced one timed argumentative CET-4 essay via the *Pigai* interface, explaining whether it is advisable to work in a state-owned business or in a joint venture. The essays were downloaded from *Pigai* platform and randomly numbered from 1 to 486, forming a small-scale learner corpus with a total of 67, 554 words.

3.4 The Rating Rubric and Procedures

We recruited two human raters to evaluate the samples. Prior to the beginning of this study, both raters had had over five years' teaching experience and had been awarded as excellent CET-4 essays raters several times. Both human raters and the two AWE systems adopted the CET-4 15-point holistic rating rubric, with the essays segmented into 5 score bands: Band 1 (1 to 3 points), Band 2 (4 to 6 points), Band 3 (7 to 9 points), Band 4 (10 to 12 points) and Band 5 (13 to15 points). This scoring rubric requires raters to conduct a comprehensive evaluation from both the content (e.g., clarity of expressing ideas and relevance to the topic) and language (e.g., accuracy, fluency and complexity in English).

To guarantee scoring accuracy and fairness, a pilot study was conducted in which 20 essays randomly selected from the corpus were rated by two human raters. Results show that the inter-rater reliability was acceptable (r = .87, p = .000). The remaining essays were divided into two halves which were evaluated independently by two human raters who negotiated and resolved the discrepancy when the score difference of one single essay exceeded 3 points. We calculated the averages of the independent scores assigned by the two raters, which were counted as the human scores. Based on both raters' rating experience and recommendations, we divided up the data set into low (those in Band 1 and 2), medium (those in Band 3 and 4) and high-quality essays (those in Band 5). The number and percentage of essays in each score band are listed in Table 1.

Table 1. The number and percentage of essays in each score band

Category	Band 1	Band 2	Band 3	Band 4	Band 5	Total
Percentage	10.70%	16.46%	30.45%	26.13%	16.26%	100%
Number	52	80	148	127	79	486

All the essays were then submitted to *iWrite* platform for obtaining *iWrite* scores. Human, *Pigai* and *iWrite* scores were input into the same excel sheet. Then, SPSS 20.0 was run to compare human-machine and machine-machine scores by calculating the maximum score difference, EPAA rate, the Pearson's *r* and Cohen's Kappa, with the latter three frequently used as evidence for the scoring validity of AWE or AES systems (e.g., Powers et al., 2002; Vantage Learning, 2002; Rudner et al., 2006; Weigle, 2010). Moreover, in the existent literature, the maximum score difference has been used less frequently, but it is supposed to reflect the difference between raters more directly (Bai & Wang, 2018) and the CET-4 scoring clearly requires the raters to control score differences.

Specifically, the maximum score difference refers to the maximum absolute value of the human-machine score difference. The EPAA rate refers to the ratio of the number of essays whose absolute value of man-machine score difference is smaller than 3 to the total number of essays. The standard for EPAA is mainly based on the CET-4 scoring rubrics (previously presented). Cohen's Kappa is a more robust measure as it corrects chance agreement (Ramineni, 2013). According to Bai and Wang (2018), the maximum score difference is negatively correlated with the scoring validity, whereas the other three show the other way around. In addition, descriptive statistics of human and machine scores were provided by SPSS 20.0 with the significance level set at 0.05.

In order to conduct an in-depth investigation into human-machine and machine-machine similarities and differences with regard to construct representation, the indices of linguistic features were gleaned manually and automatically. Statistically, all the indices were treated as independent variables and human, *Pigai* and *iWrite* scores as dependent variables to run multiple regression analyses for the establishment of corresponding scoring models for different rating methods.

3.5 Selection of Linguistic Indices

To unveil the construct representation of both systems, quantitative linguistic features of each text were collected, which fall under four categories: lexical and syntactic complexity, discourse and accuracy.

All errors of the essays were coded and counted by drawing on the classifications of Gui and Yang's (2003), Chan (2010) and Yoon and Polio (2017). For simplifying the coding process, we classified all the errors into four broad categories: mechanics, lexical, syntactic and discourse errors. Both authors coded and counted the errors in the same randomly selected 20 essays. The consistency of error coding was 94.6%, an acceptable inter-coder reliability. Then each author coded the errors of 233 essays independently. When any uncertainty emerged, both authors reached a consensus through negotiation and clarified the way to address some common problems.

Vocabprofilers (Heatley et al., 2002) was applied to analyze word frequency. Coh-Metrix (McNamara et al., 2014) and L2 Lexical Complexity Analyzer (Lu & Ai, 2015) were employed to assess word information, lexical diversity and sophistication of each sample essay. Syntactic indices relevant to syntactic diversity and complexity were computed by L2 Syntactic Complexity Analyzer (Lu, 2012). Coh-Metrix was also tapped to analyze discourse indices, including cohesion, semantic features, situation model and readability.

The information of the error types and all selected indices was set out in the Appendix.

4. Results

4.1 Response to Question 1: Human-Machine and Machine-Machine Congruence

A one-way between-groups analysis of variance was conducted to explore the mean score difference among human, *Pigai* and *iWrite* scores. As shown in Table 2, there was a statistical significance at the $p < .05$ level for the three groups: $F = 9.288$, $p = .000$. Post-hoc comparisons adopting the Turkey HSD test (see Table 3) indicated that the human mean score ($M = 8.770$, $SD = 1.950$) was significantly different from *Pigai* mean score ($M = 7.923$, $SD = 1.803$) and *iWrite* mean score ($M = 8.049$, $SD = 1.724$) with the human-*Pigai* and human-*iWrite* mean differences being 0.847 and 0.721 respectively. However, there existed no significant difference between *Pigai* and *iWrite* mean scores with a difference of only 0.126.

Table 2. Descriptive statistics for human, *Pigai* and *iWrite* scores

	N	Mean	*SD*	Minimum	Maximum	*F*	*Sig.*
Human	486	8.770	1.950	6	15		
Pigai	486	7.923	1.803	4	13	9.288	.000
iWrite	486	8.049	1.724	5	12		

Table 3. Between-group comparisons among human, *Pigai* and *iWrite* mean scores (Turkey HSD)

(I) Group	(J) Group	Mean score difference (I-J)	*Std. Error*	*Sig.*	95% Confidence interval	
					Lower bound	Upper bound
Human	*Pigai*	.847*	0.211	.000	0.35	1.34
	iWrite	.721*	0.237	.002	0.22	1.21
Pigai	human	-.847*	0.211	.000	-1.34	-0.35
	iWrite	-0.126	0.104	.822	-0.62	0.37
iWrite	human	-.721*	0.237	.002	-1.21	-0.22
	Pigai	0.126	0.104	.822	-0.37	0.62

Note. *. The mean difference is significant at the .05 level.

Table 4 reveals a significantly positive relationship between *Pigai* scores and *iWrite* scores ($r_{Pigai\text{-}iWrite}$ = .731, p < .01, *Pigai*-*iWrite* Kappa coefficient = .691, p < .01), whereas no significant correlations were found between human and machine scores ($r_{\text{human-}Pigai}$ = .158, p > .05, Human-*Pigai* Kappa coefficient = .103, p > .05; $r_{\text{human-}iWrite}$ = .122, p > .05; Human-*iWrite* Kappa coefficient = .118, p > .05).

Table 4. Correlations among human, *Pigai* and *iWrite* scores

	N	Pearson's correlation coefficient		Cohen's Kappa	
		r	*Sig.*	*value*	*Sig.*
Human-*Pigai*	486	.158	.053	.103	.067
Human-*iWrite*	486	.122	.136	.118	.756
Pigai-*iWrite*	486	.731	.000	.691	.000

Table 5 indicates the maximum score differences between human scores and machine scores were alarmingly high, 9 points and 7 points for human-*Pigai* and human-*iWrite* pairs respectively. What deserves due attention is that all the essays were scored on a 15-point scale, so a large discrepancy was found between human raters and the two systems. By contrast, the maximum score difference between machine scores was comparatively small, with only 3.4 points.

Table 5. Maximum score difference between different pairs

Pairs	Human-*Pigai*	Human-*iWrite*	*Pigai*-*iWrite*
Maximum score difference	9	7	3.4

The EPAA rates, displayed in Table 6, between human scores and *Pigai* and *iWrite* scores were quite close, 74.1% and 77.2% respectively. Human raters agreed perfectly with *Pigai* 7.1% of the time, and with *iWrite* 2% of the time. Two systems agreed with each other most of the time (EPAA rate = 97.5%) and only assigned 12 discrepant scores.

Table 6. Between-group agreement rates for human, *Pigai* and *iWrite* scores

	Human-*Pigai*	Human-*iWrite*	*Pigai*-*iWrite*
Exact agreement rate	7.1% (*n*=35)	2% (*n*=10)	0.8% (*n*=4)
0<score difference<=1	36% (*n*=175)	31.3% (*n*=152)	57.4% (*n*=279)
1<score difference<=2	22.6% (*n*=110)	28% (*n*=136)	31.3% (*n*=152)
2<score difference<3	8.2% (*n*=40)	15.8% (*n*=77)	8.0% (*n*=39)
EPAA rate	74.1% (*n*=360)	77.2% (*n*=375)	97.5% (*n*=474)
Score difference>=3	25.9% (*n*=126)	22.8% (*n*=111)	2.5% (*n*=12)

4.2 Response to Question 2: The Inconsistently-Graded Essay Type by Human and AWE Systems

As revealed in Table 7, the essays with high human scores (Band 5) tended to be assigned significantly lower scores by both AWE systems with the human-machine mean score difference exceeding 4 points; furthermore, human and machine scores agreed the least for this group with EPAA rates less than 40%. Human and machine scores highly agreed with each other for M (Band 3 and 4) and L (Band 1 and 2) group essays. One-way ANOVA analysis showed that there were significant differences among the means of the three groups' score differences (p = .000). The post hoc Turkey's test shows no significant difference between the mean score differences of essays in L and M groups for all three pairs (i.e., human-*Pigai*, human-*iWrite and Pigai-iWrite*), no significant difference between the mean score differences for *Pigai-iWrite* pair, but a significant one for the other two pairs in group H essays (p < .01).

Table 7. Agreement and mean score difference among low-, medium- and high-score essays

	L (n = 132)	M (n = 275)	H (n = 79)
Human-*Pigai EPAA*	92.85%	81.37%	37.5%
Human-*Pigai* mean score difference	1.43[a]	1.88[a]	4.11[b]
Human-*iWrite EPAA*	91.8%	80.3%	31.25%
Human-*iWrite* mean score difference	1.5[a]	2[a]	4.06[b]
Pigai-iWrite EPAA	97.3%	96.7%	100%
Pigai-iWrite mean score difference	0.92[a]	1.06[a]	1.41[a]

Note. One-way ANOVA was used to compare the absolute values of the mean score differences of essays in low-, medium- and high-quality groups, and Turkey method was used to carry out multiple comparisons afterwards. The same superscript letters (e.g., a, b, c) on the mean demonstrate no significant difference between groups, while the different letter indicates a significant difference with other groups (p < .01).

Table 8 displays the percentages of essays with discrepant human-machine scores at different quality levels. It also reveals that high-human-score essays might be prone to be assigned low scores by both AWE systems whose discrepancy levels reached 68.4%, much higher than those for essays in the other two groups. From Table 7 and Table 8, a conclusion could be drawn that the essays considered to be of high quality by human raters would be largely considered to be of poor quality by AWE systems.

Table 8. Percentages of essays with discrepant human-machine scores

	L (n = 132)	M (n = 275)	H (n = 79)
Human-*Pigai*	13.6% (n = 18)	29.5% (n = 81)	68.4% (n = 54)
Human-*iWrite*	8.3% (n = 11)	28.0% (n = 77)	62.0% (n = 49)

4.3 Response to Question 3: Impacts of Quantitative Features on Human and Machine Scores

A stepwise regression analysis was run to figure out the relationship between the dependent variable (human scores) and the independent variables (error, syntactic, lexical and discourse indices). Among the 9 obtained regression models, the ninth with the best goodness of fit was chosen as the scoring model of human scores. Table 9 shows that nine indices entered this model, with all of them significantly contributing to the prediction (p < .05) and no co-linearity problem among them. Nine indices accounted for 48.1% of the variance (Adjusted R^2 = .481, F (9, 139) = 16.243, p = .000). The regression formulation is: Human scores = -31.930+0.103×WRDFAMc+0.054×CNCADC-0.129×Syntactic-0.017×WRDADJ-0.022×WRDADV-4.479×WRDFRQa-3.054×LV+0.033×CNCTempx+0.086×SLEXTYPES.

Table 9. Multiple linear regression of human scores: Important statistics (n = 486)

Variables	R	R^2	Adjusted R^2	F	*Beta*	*t*	Sig.	Tolerance	VIF
Human scores	.716	.513	.481	16.243 (9, 139)					
Constant					-31.930	-3.776	.000		
WRDFAMc					.103	6.337	.000	.662	1.511
CNCADC					.054	4.764	.000	.950	1.052
Syntactic					-.129	-3.891	.000	.856	1.168
WRDADJ					-.017	-4.385	.000	.786	1.272
WRDADV					-.022	-4.498	.000	.858	1.166
WRDFRQa					-4.479	-3.076	.003	.665	1.504
LV					-3.054	-3.003	.003	.822	1.216
CNCTempx					.033	2.530	.013	.915	1.093
SLEXTYPES					.086	2.204	.029	.892	1.121

In a similar vein, 15 models for *Pigai* scores were obtained, and the 12th was selected as the scoring model. Table 10 shows that 12 indices entered this model, with all of them significantly contributing to the prediction ($p < .05$) and no co-linearity problem among them. All indices predicted 73.3% of the score variance (Adjusted R^2=.733, $F(12,136) = 30.225$, $p = .000$). The regression formulation is: *Pigai* scores = -0.757+0.257×MLC-5.142×ERROR rate+1.264×CTTR-0.017×WRDPRO-0.190×Mechanical+0.799×C/S-3.893×SMCAUSlsa+0.433×WRDHYPn-8.822×K2 (%)+0.931×WRDHYPv-1.930×LV+3.652×VS1.

Table 10. Multiple linear regression of *Pigai* scores: Important statistics (n = 486)

Variables	R	R^2	Adjusted R^2	F	*Beta*	*t*	*Sig.*	Tolerance	*VIF*
Pigai scores	.869	.755	.733	30.225 (12, 136)					
Constant					-.757	-.486	.628		
MLC					.257	4.566	.000	.548	1.826
ERROR rate					-5.142	-2.119	.036	.472	2.119
CTTR					1.264	5.999	.000	.680	1.471
WRDPRO					-.017	-8.074	.000	.461	2.168
Mechanical					-.190	-5.984	.000	.444	2.251
C/S					.799	3.888	.000	.773	1.294
SMCAUSlsa					-3.893	-2.136	.034	.931	1.074
WRDHYPn					.433	3.214	.002	.876	1.142
K2(%)					-8.822	-3.137	.002	.648	1.544
WRDHYPv					.931	2.498	.014	.941	1.063
LV					-1.930	-2.534	.012	.648	1.543
VS1					3.652	2.436	.016	.891	1.122

10 models for *iWrite* scores were obtained, and the 10th was selected as the scoring model. Table 10 shows that 10 indices entered this model, with all of them significantly contributing to the prediction ($p < .05$) and no co-linearity problem among them. All indices explained 77.6% of the variance (Adjusted $R^2 = .776$, $F(10,138) = 52.228$, $p = .000$).The regression formulation is: *iWrite* scores = 9.790-0.01×WRDPRO-0.157×Mechanical-8.515×K2(%)+0.103×MLC+1.315×LSAPP1+0.023×WRDPRP1s-0.007×CNCALL-0.058×VP-1.927×LSAPP1d+0.15×WORDTYPES.

Table 11. Multiple linear regression of *iWrite* scores: Important statistics (n = 486)

Variables	*R*	R^2	Adjusted R^2	*F*	*Beta*	*t*	*Sig*.	Tolerance	*VIF*
iWrite scores	.889	.791	.776	52.258 (10, 138)					
Constant					9.790	12.833	.000		
WRDPRO					-.010	-4.462	.000	.270	3.708
Mechanical					-.157	-7.840	.000	.845	1.184
K2 (%)					-8.515	-3.209	.002	.551	1.815
MLC					.103	1.885	.061	.441	2.267
LSAPP1					1.351	2.939	.004	.531	1.884
WRDPRP1s					.023	3.433	.001	.887	1.128
CNCALL					-.007	-2.413	.017	.483	2.069
VP					-.058	-3.128	.002	.268	3.738
LSAPP1d					-1.927	-2.262	.025	.937	1.068
WORDTYPES					.015	2.032	.044	.506	1.977

5. Discussion

5.1 Agreement among Human, Pigai and iWrite Scores

To summarize, *Pigai* and *iWrite* agreed more with each other than with human raters in rating the 486 essays. This is firstly evidenced by the high *Pigai-iWrite* EPAA rate of 97.5% and the human-machine EPAA rates lower than 80%. Burstein et al. (2004) pointed out that AES systems could be seen as reliable measurement tools only when the human-machine EPAA figures could reach the baseline of 75%–80%. In this sense, only *iWrite* just met the basic requirement despite the high machine-machine agreement. It should be noted that the EPAA rate might be misleading due to its sensitivity to size of research samples (Yang et al., 2002), but how the sample size would exert an influence on the research results remains to be examined. In terms of Pearson's r and Cohen's Kappa, scores assigned by both systems agreed much (r = .731, p = .000; Kappa coefficient = .691, p = .000), whereas machine scores were not significantly correlated with human scores (p > .05) with Pearson's r (.158 for *Pigai* and .122 for *iWrite*) far under the threshold of 0.7 level required in this line of research (Ramineni & Williamson, 2013), and the Kappa coefficients were equally unsatisfactory (.103 for *Pigai* and .118 for *iWrite*). Last, the *Pigai-iWrite* maximum score difference was much lower than human-machine ones. From these data, a tentative inference can be made here that the scoring mechanisms of *Pigai* and *iWrite* might follow some similar patterns, especially with regard to their scoring methods and the valued writing constructs, while both systems are different from human raters in these aspects. These machine-machine and human-machine similarities and differences will be discussed in the third section of the discussion part.

In terms of human-machine comparison, EPAA rates between human and systems like PEG, *e-rater*, IntelliMetric or *Criterion* were reported to be much higher than those found in the present study. For example, *e-rater*'s EPAA rates averaged 90% or above (Valenti et al., 2003) and the exact agreement rates ranged from 48% to 58% (Wang & Brown, 2008), and even approached 80% (Shermis et al., 2008). Divergent rating scales might very well account for such a discrepancy between the results obtained by the present study and those by previous studies (Ramineni & Williamson, 2013). To be sure, human-machine score agreement based on a 3-point rating scale is bound to be higher than that based on a 10-point one. When a human rater assigns an essay 2 points, the system may assign 1 point and 3 points. Both scores are within one-point discrepancy with the human score, and thus they are adjacently agreeing with the human score, which may inflate the agreement rate. Previous studies mostly employed a 4-point or 6-point scale, likely to elevate human-machine agreement rates. Shermis and Hamner (2012) also expounded that the EPAA rate of 100% could be obtained for a 3-point rating scale, 99%, 55% and 49% for 6-point, 12-point and 30-point rating scales. In this sense, the adoption of a 15-point scale in the present study may partially account for the relatively low human-machine agreement rate. But this explanation still needs to be validated in future research.

Equally, a cornucopia of studies reported on quite high human-machine correlation coefficients which mostly surpassed or approached the baseline level (Burstein & Chodorow, 1999; Shermis et al., 2002; Vantage Learning, 2003; Weigle, 2010; Ramineni, 2013). Again, the rating scales serve as a possible factor for the research results, as Shermis and Hamner (2013) implied a coefficient about 0.75 for a 4-point scale, 0.72 for a 12-point scale and 0.61 for a 30-point scale.

5.2 AWE Systems' Proneness to Assign Low Scores to High-human-Score Essays

Another finding is that both systems agree more with human raters when evaluating the low- and medium-quality

essays (according to human judgment) with the EPAA rates exceeding 80%, but less with human raters when assessing the high-quality essays with the agreement rates lower than 40%. Several studies reported on the unreliability of AES systems to assess high-quality essays. Burstein et al. (1998) investigated the agreement between human scorers and e-rater (with the sample essays scored on a 6-point scale) and found that the greatest discrepancy lay in band 5 and 6 essays. Li et al. (2014) also discerned an analogous flaw in *Criterion*. Somewhat differently, Ge and Chen (2009) pointed out the objectivity of low machine scores and the inappropriateness of the moderate and high machine scores (high scores in particular), but provided no tangible proofs or detailed explanations to straighten out the fallibility of the systems in evaluating essays.

The research findings of the present study were a perfect echo of what was revealed by Bai and Wang (2018), which pointed out *Pigai*'s taking the high-human-score essays as low-quality ones. *Pigai* and *iWrite* can, as pointed out by Bai and Wang (2018), accurately score the essays with low human scores, presumably and in large measure, owing to the poor language quality of this type of essay. Machines are in a good position to assign objective scores based on the superficial quantifiable features or language errors (Bai & Wang, 2018). Likewise, when evaluating low-quality essays, human scorers would still put emphasis on the quantifiable features, although these features may have nothing to do with what makes a good essay (Condon, 2013), since an excellent essay may be characterized by the diction, the clear-cut structure, the originality of ideas, the reasonable demonstration of an argument or the mixture of all these features (Bai, 2011). Chances are that due to low English proficiency some students involved in this study tended to select high-frequency vocabulary or common expressions and sentence patterns in their essays to reduce errors. Despite lack of sophisticated words and sentences, their essays may be well-organized, original in ideas, abundant in rhetorical use, etc., and so will be favored by human raters. However, this speculation remains to be addressed by complementary qualitative evidence with regard to all these aspects not resolved in the present study.

5.3 Differing Impacts of Quantitative Linguistic Features on Human and Machine Scoring

The results demonstrate that the two systems might have valued different writing constructs, since different variables remained in the regression equations for human and machine scores and explained 48.1%, 73.3% and 77.6% of the human, *Pigai* and *iWrite* scores respectively. This study produced mixed results with Bai and Wang (2018) which showed that the quantitative features entering the regression equations accounted for over 65% of the machine score variance but less than 25% of the human score variance. The reason for such a difference may lie in the fact that Bai and Wang (2018) selected far fewer quantitative features. We assume that more quantitative features are likely to inflate the explanatory power. But on the whole, both studies found greater impacts of the quantitative features on machine scores than on human scores, and showed the heavy reliance of *Pigai* and *iWrite* on counting the quantifiable linguistic items. However, it was also found that different variables entered the regression models of *Pigai* and *iWrite*, suggesting both systems might look at different aspects when scoring essays despite the similar statistical results. This difference may result from the establishing process of both systems in which the developers may have selected different quantitative linguistic features to train the scoring models. But since there is no literature reporting on the process of both systems, we cannot have a clear picture of the developers' selection of linguistic features.

The divergence in the scoring equations of the human and machine scores could, in large measure, show that the human scorers and the systems put emphasis on different features of the essays, i.e., different construct representations. For example, mechanical error entered the scoring models of both systems, while syntactic error only remained in the human scoring equation. The former includes spelling, word building, capitalization and punctuation errors, all of which can be easily identified by the systems (Wan, 2005; Jiang et al., 2011; Yang, 2013) or even by any word processing software, so it comes as no surprise that the AWE systems can also do a good job. The fundamental problem is the mechanical way of the systems to identify and judge errors (Shi, 2012; Yang, 2013). To put it another way, it is hard for the systems to judge more sophisticated syntactic errors. For instance, when treating sentence fragments like 'Even find it difficult in grabbing the ball', human scorers can easily find that the subject of the sentence is missing in this sentence, while machine is likely to take the adverb 'even' for the subject.

This difference in the quantitative features' predictive powers on human and machine scoring can be explained by the divergence in the human-machine scoring processes. In human scoring, although formal factors such as essay length, lexical richness, syntactic complexity and discourse coherence will affect the score, the specific scoring is a very complicated cognitive process. Wolfe's (1997) think-aloud study showed that teachers first read articles and formed text schemata in line with their own background, viewpoints, writing knowledge, etc. These schemata were not a copy of the original text, but were integrated with teachers' own understanding and judgement. During the reading process, teachers would monitor the content and characteristics of an essay. Finishing reading, they would

re-examine the text schemata in compliance with the scoring criteria to determine the extent to which the specific scoring criteria were consistent and finally scored the essay. As far as the specific scoring process is concerned, Wolfe (2005) found that experienced teachers generally adopted a top-down cognitive model, i.e., to judge the essay as a whole. For example, for the tense problems, instead of pointing out the errors one by one, they would assign the score after reading the whole essay. But novice teachers would often make judgments before reading the essay, and then adjusted their judgment during the reading process. The reason why such a difference existed was mainly that seasoned teachers had established a complex information processing network, which enabled them to store a large amount of information in the reading process and finally allowed them to carry out a comprehensive processing. But the novices often did not have such an ability. In short, experienced teachers would make a comprehensive evaluation of the essay, and novices tended to pay too much attention to the details.

Contrarily, an essay is just an accumulation of words for the machine. Essay evaluation resembles a simple stimulus-response process, and the machine can only respond to various stimuli already set in the program (Ericsson, 2006), which is completely different from the construction process of text schemata in human evaluation. In addition, like novice teachers, the systems focus only on details, such as the number of conjunctions, the proportion of complex words, the average length of sentences, etc., quantify all indicators of essay quality, and then assign a score according to the weights of different aspects. This is a bottom-up approach, which is entirely opposite to the judging process of experienced teachers. Therefore, it is not difficult to explain why quantitative linguistic features with sufficient explanatory power in machine scoring have little effect in human scoring.

6. Conclusion

This study evaluated the scoring validity of two commercially available Chinese AWE systems by sampling 486 timed essays produced by second-year non-English majors, and the research findings show a barely satisfactory scoring performance of the systems. Based on these findings, we provided in-depth explanations for the underlying causes.

It is important to show clearly several limitations of the present study. First, this study did not deeply analyze the language and content of the essays most prone to be inconsistently evaluated by human raters and AWE systems. Second, we adopted a quantitative approach and did not address the content of the essays, and so the picture of the writing constructs could be partially explained. In the future, researchers in this field can employ a mixed research methodology (both qualitative and quantitative) and address the scoring validity more comprehensively. Despite these limitations, it is worth affirming that this study has played a warning role for the improvement of Chinese AWE systems, the integration of AWE systems into L2 English writing instruction, and the integration of machine scores into students' final scores in the Chinese teaching settings. It is our belief that this field needs to draw the attention of more independent researchers and users.

References

Bai, L. F. (2011). The relationship between lexical competence and the quality of L2 writing: an overview. *Foreign Language Education in China*, *4*(4), 3–10.

Bai, L. F., & Hu, G. W. (2017). In the face of fallible AWE feedback: how do students respond? *Educational Psychology*, *37*(1), 67–81. https://doi.org/10.1080/01443410.2016.1223275

Bai, L. F., & Wang, J. (2018). Difference between human and machine scoring and its underlying causes. *Foreign Language Testing and Teaching*, *3*, 44–54.

Bai, L. F., & Wang, J. (2019). A critical review of the effectiveness of English AWE feedback over the past 20 years. *Foreign Languages Research*, *36*(1), 65–71, 88.

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

Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The criterion online writing service. *AI Magazine*, *25*(3), 27–36. https://dl.acm.org/doi/10.5555/1045744.1045749

Burstein, J., Kukich, K., Braden-Harder, L., Chodorow, M., Hua, S., & Kaplan, B. (1998). *Computer analysis of essay content for automatic score prediction: A prototype automated scoring system for GMAT analytical writing assessment*. (ETS Research Report RR-98-15). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1998.tb01764.x

Chan, A. Y. (2010). Toward a taxonomy of written errors: Investigation into the written errors of Hong Kong Cantonese ESL learners. *Tesol Quarterly*, *44*(2), 295–319. https://doi.org/10.5054/tq.2010.219941

Chen, C. F. E., & Cheng, W. Y. E. C. (2008). Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes. *Language Learning & Technology*, *12*(2), 94–112. https://dx.doi.org/10125/44145

Chen, X. X., & Ge, S. L. (2008). A review of automatic essay scoring. *Journal of PLA University of Foreign Languages*, *31*(5), 78–83. Retrieved from https://peerj.com/articles/cs-208

Cheville, J. (2004). Automated scoring technologies and the rising influence of error. *The English Journal*, *93*(4), 47–52. https://doi.org/10.2307/4128980

Choi, I.-C. (2014). Efficacy of an ICALL tutoring system and process-oriented corrective feedback. *Computer Assisted Language Learning*, *29*(2), 334–364. https://doi.org/10.1080/09588221.2014.960941

Condon, W. (2013). Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings? *Assessing Writing*, *18*(1), 100–108. https://doi.org/10.1016/j.asw.2012.11.001

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. Retrieved from https://files.eric.ed.gov/fulltext/EJ843855.pdf

Elliot, N., & Williamson, D. M. (2013). Assessing writing special issue: Assessing writing with automated essay scoring. *Assessing Writing*, *18*(1), 1–6. https://doi.org/10.1016/j.asw.2012.11.002

Ericsson, P. F. (2006). The meaning of meaning: Is a paragraph more than an equation. In P. F. Ericsson & R. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 28–37). Logan: All USU Press Publications. https://doi.org/10.2307/j.ctt4cgq0p.5

Gao, J., Li, X., Gu, P., & Liu, Z. (2020). An evaluation of china's automated scoring system bingo English. *International Journal of English Linguistics*, *10*(6), 30. https://doi.org/10.5539/ijel.v10n6p30

Ge, S. L., & Chen, X. X. (2009). The key problems and solutions in automated essay scoring for college English teaching in China. *Shandong Foreign Language Teaching Journal*, *30*(3), 21–26.

Grimes, D., & Warschauer, M. (2008). Learning with laptops: A multi-method case study. *Journal of Educational Computing Research*, *38*(3), 305–332. https://doi.org/10.2190/EC.38.3.d

Gu, C. H., & Wang, L. (2012). An empirical study of college writing teaching based on Juku correcting network. *Journal of Yangzhou University* (Higher Education Study Edition), *16*(4), 92–96.

Gui, S. C., & Yang, H. Z. (2003). *Chinese Learner English Corpus*. Shanghai: Shanghai Foreign Language Education Press.

He, X. L. (2013). Reliability and validity of the assessment of the *Pigaiwang* on college students' writings. *Modern Educational Technology*, *23*(5), 64–67.

Heatley, A., Nation, I. S. P., & Coxhead, A. (2002). *Range and Frequency programs*.

Jiang, X. Q., Cai, J., & Tang, J. L. (2011). Impacts of an automated essay scoring tool on the development of writing proficiency of Chinese college EFL learners. *Shandong Foreign Language Teaching Journal*, *32*(6), 36–43.

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64), Westport, CT: American Council on Education and Praeger.

Lee, C., Wong, K. C. K., Cheung, W. K., & Lee, F. S. L. (2009). Web-based essay critiquing system and EFL students' writing: a quantitative and qualitative investigation. *Computer Assisted Language Learning*, *22*(1), 57–72. https://doi.org/10.1080/09588220802613807

Li, J. H. (2009). *Using latent semantic analysis for automated essay scoring in the Chinese EFL context*. Guang Zhou: Guangdong University of Foreign Studies.

Li, Y. L., & Tian, X. C. (2018). An empirical research into the reliability of *iWrite* 2.0. *Modern Educational Technology*, *28*(2), 75–80.

Li, Z., Link, S., Ma, H., Yang, H., & Hegelheimer, V. (2014). The role of automated writing evaluation holistic scores in the ESL classroom. *System*, *06*(44), 66–78. https://doi.org/10.1016/j.system.2014.02.007

Liang, M. C. (2011). *Constructing a model for the computer-assisted scoring of Chinese EFL learners' argumentative essays*. Beijing: Foreign Language Teaching and Research Press.

Liang, M. C., & Wen, Q. F. (2007). A review of foreign automatic essay scoring systems and their implications. *Computer-Assisted Foreign Language Education*, *5*, 18–24.

Lu, X. F. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, *96*(2), 190–208. https://doi.org/10.1111/j.1540-4781.2011.01232.x

Lu, X. F., & Ai, H. Y. (2015). Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. *Journal of Second Language Writing*, *29*, 16–27. https://doi.org/10.1016/j.jslw.2015.06.003

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–104). New York, NY: American Council on education and Macmillan. https://doi.org/10.1017/CBO9780511894664

Powers, D. E., Burstein, J. C., Chodorow, M. S., Fowles, M. E., & Kukich, K. (2002). Comparing the validity of automated and human scoring of essays. *Educational Computing Research*, *26*(4), 407–425. https://doi.org/10.2190/CX92-7WKV-N7WC-JL0A

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

Ramineni, C. (2013). Validating automated essay scoring for online writing placement. *Assessing Writing*, *18*(1), 40–61. https://doi.org/10.1016/j.asw.2012.10.005

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

Ranalli, J. (2018). Automated written corrective feedback: how well can students make use of it? *Computer Assisted Language Learning*. https://doi.org/10.1080/09588221.2018.1428994

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. Retrieved from https://files.eric.ed.gov/fulltext/EJ843850.pdf

Sarré, C., Grosbois, M., & Brudermann, C. (2019). Fostering accuracy in L2 writing: impact of different types of corrective feedback in an experimental blended learning EFL course. *Computer Assisted Language Learning*. https://doi.org/10.1080/09588221.2019.1635164

Shermis, M. D., & Hamner, B. (2013). Contrasting State-of-the-Art Automated Scoring of Essays. In M. D. Shermis & J. Burstein (Eds.), *Handbook of Automated Essay Evaluation* (pp. 213–246). New York, NY: Routledge. https://doi.org/10.4324/9780203122761

Shermis, M. D., Koch, C. M., Page, E. B., Keith, T. Z., & Harrington, S. (2002). Trait ratings for automated essay grading. *Educational and Psychological Measurement*, *62*(1), 5–18. https://doi.org/10.1177/0013164402062001001

Shermis, M. D., Shneyderman, A., & Attali, Y. (2008). How important is content in the ratings of essay assessments? *Assessment in Education: Principles, Policy and Practice*, *15*(1), 91–105. https://doi.org/10.1080/09695940701876219

Shi, X. L. (2012). A tentative study on the validity of online automated essay scoring used in the teaching of EFL writing—exemplified by http//www.*Pigai*.org. *Modern Educational Technology*, *22*(10), 67–71.

Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. *Assessing Writing*, *19*(1), 51–65. https://doi.org/10.1016/j.asw.2013.11.007

Valenti, S., Neri, F., & Cucchiarelli, A. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education*, *2*, 319–330. https://doi.org/10.28945/331

Vantage Learning. (2002). *A study of IntelliMetric® scoring for responses written in Bahasa Malay (No. RB-735)*. Newtown, PA: Vantage Learning.

Vantage Learning. (2003). *How does IntelliMetric® score essay responses (No. RB-929)*. Newtown, PA: Vantage Learning.

Wan, P. J. (2005). Research report on testing college English writing with electronic software assessment system. *Media in Foreign Language Instruction*, *3*, 11–13, 31.

Wang, H. J. (2016). An empirical research into scoring validity of AES. *Journal of Zhejiang University of Technology (Social Science)*, *15*(1), 89–93.

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

Wang, J., & Zhang, T. Y. (2020). A study of the relationship between the textual quantitative indices of L2 writing and machine scores. *Foreign Language Testing and Teaching (Quarterly)*, *39*(3), 12–20. https://doi.org/10.5070/L212245911

Wang, Y. Y. (2012). The scoring validity of *Write On* automated essay scoring system. *Theory and Practice of Contemporary Education*, *4*(12), 139–142.

Weigle, S. C. (2010). Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability. *Language Testing*, *27*(3), 335–353. https://doi.org/10.1177/0265532210364406

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. London: Palgrave Macmillan. https://doi.org/10.1057/9780230514577

Wolfe, E. W. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assessing Writing*, *4*(1), 83–106. https://doi.org/10.1016/S1075-2935(97)80006-2

Wolfe, E. W. (2005). Uncovering rater's cognitive processing and focus using think-aloud protocols. *Journal of Writing Assessment*, *2*(1), 37–56.

Xu, Y. (2018). Investigating the Validity of Using Automated Writing Evaluation in EFL Writing Assessment. In T. Hao, W. Chen, H. Xie, W. Nadee & R. Lau (Eds.), *Emerging Technologies for Education*. SETE 2018. Lecture Notes in Computer Science, vol 11284. Springer, Cham. https://doi.org/10.1007/978-3-030-03580-8_14

Yang, L. (2013). On the application of AWE system in high-level students' EFL writing learning. *Modern Educational Technology*, *23*(5), 73–77.

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*(4), 391–412. https://doi.org/10.1207/S15324818AME1504_04

Yoon, H. J., & Polio, C. (2017). The linguistic development of students of English as a second language in two written genres. *Tesol Quarterly*, *51*(2), 275–301. https://doi.org/10.1002/tesq.296

Zhang, L. (2017). A study on the evaluation, generalization and extrapolation of automated writing evaluation system. *Foreign Language and Translation, 24*(3), 64–71, 98.

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

Ziegler, N., & Mackey, A. (2017). Interactional feedback in synchronous computer-mediated communication: A review of the state of the art. In H. Nassaji & E. Kartchava (Eds.), *Corrective feedback in second language teaching and learning: Research, theory, applications, implications* (pp. 80–94). New York, NY: Routledge. https://doi.org/10.4324/9781315621432-7-

Appendix A

Quantitative Essay Features at Four Levels

I. Accuracy Level

Sub-categories	Specific items	Tool (s) used
Mechanics	Capitalization, punctuation, misspelling	Manual checking and classification
Lexical	Word class, plural forms, word choice, articles, lexical collocation, gender and case of pronouns, etc.	
Syntactic	Word order, fragments, run-ons, missing parts, redundancy, subordinate clause, subject-predicate agreement, tense and voice, non-finite verbs, etc.	
Discourse	Connectives, reference, etc.	

II. Lexical Complexity

Sub-categories	Specific items	Tool (s) used
Word Types & tokens	WORDTYPES, SWORDTYPES, LEXTYPES, SLEXTYPES, WORDTOKENS, SWORDTOKENS, LEXTOKENS, SLEXTOKENS	L2 Lexical Complexity Analyzer
Lexical sophistication	Lexical sophistication-I; Lexical sophistication-I; Verb sophistication-I; Verb sophistication-II; Corrected VS1	
Lexical variation (NDW)	Number of different words; NDW (first 50 words); NDW(expected random 50); NDW (expected sequence 50)	
Lexical variation (TTR)	Type/token ratio; Mean segmental TTR; Corrected TTR; Root TTR; Biogarithmic TTR; Uber index	
Lexical variation (Verb diversity)	Verb variation; Squared VV1; Corrected VV1	
Lexical variation (Lexical word diversity)	Lexical word variation; verb variation-II; noun variation; adjective variation; adverb variation; modifier variation	
WRDNOUN	Noun incidence	Coh-Metrix 3.0
WRDVERB	Verb incidence	
WRDADJ	Adjective incidence	
WRDADV	Adverb incidence	
WRDPRO	Pronoun incidence	
WRDPRP1s	First person singular pronoun incidence	
WRDPRP1p	First person plural pronoun incidence	
WRDPRP2	Second person pronoun incidence	
WRDPRP3s	Third person singular pronoun incidence	
WRDPRP3p	Third person plural pronoun incidence	
WRDFRQc	CELEX word frequency for content words, mean	
WRDFRQa	CELEX Log frequency for all words, mean	
WRDFRQmc	CELEX Log minimum frequency for content words, mean	
WRDAOAc	Age of acquisition for content words, mean	
WRDFAMc	Familiarity for content words, mean	
WRDCNCc	Concreteness for content words, mean	
WRDIMGc	Imagability for content words, mean	
WRDMEAc	Meaningfulness, Colorado norms, content words, mean	
WRDPOLc	Polysemy for content words, mean	
WRDHYPn	Hypernymy for nouns, mean	
WRDHYPv	Hypernymy for verbs, mean	
WRDHYPnv	Hypernymy for nouns and verbs, mean	
SMCAUSv	Causal verb incidence	
SMCAUSvp	Causal verbs and causal particles incidence	
SMINTEp	Intentional verbs incidence	
SMCAUSr	Ratio of casual particles to causal verbs	
SMINTEr	Ratio of intentional particles to intentional verbs	
SMCAUSlsa	LSA verb overlap	
SMCAUSwn	WordNet verb overlap	
SMTEMP	Temporal cohesion, tense and aspect repetition, mean	

III. Syntactic Complexity

Sub-categories	Specific items	Tool (s) used
Syntactic structures	Word count; sentence; verb phrase; clause; T-unit; dependent clause; complex T-unit; Coordinate phrase; complex nominal	L2 Syntactic Complexity Analyzer
Syntactic complexity indices	Mean length of sentence; mean length of T-unit; mean length of clause; clause per sentence; verb phrase per T-unit; clause per T-unit; dependent clause per clause; dependent clause per T-unit; T-unit per sentence; complex T-unit ratio; coordinate phrase per T-unit; coordinate phrase per clause; complex nominal per T-unit; complex nominal per clause	

IV. Discourse Indices

Sub-categories	Specific items	Tool (s) used
CRFNO1	Noun overlap, adjacent sentences, binary, mean	Coh-Metrix 3.0
CRFAO1	Argument overlap, adjacent sentences, binary, mean	
CRFSO1	Stem overlap, adjacent sentences, binary, mean	
CRFNOa	Noun overlap, all sentences, binary, mean	
CRFAOa	Argument overlap, all sentences, binary, mean	
CRFSOa	Stem overlap, all sentences, binary, mean	
CRFCWO1	Content word overlap, adjacent sentences, proportional, mean	
CRFCWO1d	Content word overlap, adjacent sentences, proportional, standard deviation	
CRFCWOa	Content word overlap, all sentences, proportional, mean	
CRFCWOad	Content word overlap, all sentences, proportional, standard deviation	
LSASS1	LSA overlap, adjacent sentences, mean	
LSASS1d	LSA overlap, adjacent sentences, standard deviation	
LSASSp	LSA overlap, all sentences in paragraph, mean	
LSASSpd	LSA overlap, all sentences in paragraph, standard deviation	
LSAPP1	LSA overlap, adjacent paragraphs, mean	
LSAPP1d	LSA overlap, adjacent paragraphs, standard deviation	
LSAGN	LSA given/new, sentences, mean	
LSAGNd	LSA given/new, sentences, standard deviation	
CNCAll	All connectives incidence	
CNCCaus	Causal connectives incidence	
CNCLogic	Logical connectives incidence	
CNCADC	Adversative and contrastive connectives incidence	
CNCTemp	Temporal connectives incidence	
CNCTempx	Expanded temporal connectives incidence	
CNCAdd	Additive connectives incidence	
CNCPos	Positive connectives incidence	
CNCNeg	Negative connectives incidence	
RDFRE	Flesch Reading Ease	
RDFKGL	Flesch-Kincaid Grade level	
RDL2	Coh-Metrix L2 Readability	