Investigating the validity of using automated writing evaluation in EFL writing assessment

Ying Xu

July 3, 2018

Investigating the validity of using automated writing evaluation in EFL writing assessment[1]

Ying Xu[1]

[1] School of Foreign Languages, South China University of Technology, Guangzhou 510641, China

xuying@scut.edu.cn

Abstract. This study aims to follow an argument-based approach to validation of using automated essay evaluation (AWE) system with the example of *Pigai*, a Chinese AWE program, in English as a Foreign Language (EFL) writing assessment in China. First, an interpretive argument was developed for its use in the course of College English. Second, three sub-studies were conducted to seek evidence of claims related to score evaluation, score generalization, score explanation, score extrapolation and feedback utilization. Major findings are: (1) *Pigai* yields scores that are accurate indicators of the quality of a test performance sample; (2) its scores are consistent across tasks in the same form; (3) its scoring features represent the construct of interest to some extent, yet problems of construct under-representation and construct-irrelevant features still exist; (4) its scores are consistent with teachers' judgments of students' writing ability; (5) its feedback has a positive impact on students' development of writing ability, but to some extent. These results reveal that AWE can only be used as a supplement to human evaluation, but can never replace the latter.

Keywords: *Pigai,* Automated essay evaluation, Writing assessments.

1 Introduction

With the technology boom in the recent years, automated writing evaluation (AWE) has developed rapidly because it can provide immediate feedback on students' essays to a large EFL writing class. However, the traditional approach to AWE validity mainly focuses on the system's psychometric properties while classroom users' responses and perceptions are neglected [1]. The typical way to validating an AWE system is to calculate the correlation between machine scores and human scores. To date, almost all vendor-sponsored research claim a high correlation coefficient for systems such as *PEG*, *IEA*, and *E-rater* [2]. As a response to Warschauer's call for more independent research, a handful of researchers try to introduce the latest development in test validity to the field of AWE. For example, Xi [3] raised ten

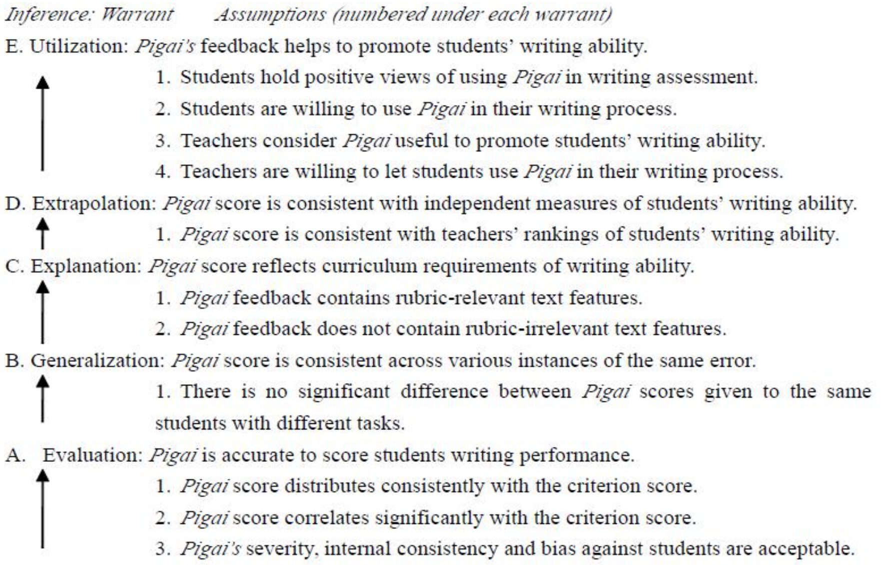
This work was funded by the Ministry of Education, P. R. China (Number 17YJC740102).

fundamental questions for automated scoring systems. On its basis, a framework for evaluation and use of automated scoring was built [4], which clarifies inferences in terms of explanation, evaluation, extrapolation, generalization, and utilization within an argument-based validity. Given that this framework requires various categories of data, empirical studies which have adopted the framework are scanty.

In China, *Pigai* (www.pigai.org) was developed specifically to assess Chinese EFL learners' writing [5]. At the time of writing, it was reported that *Pigai* was used by over 1,000 universities in China. The scoring engine, calibrated against a large corpus of human-scored essays, can generate a score and feedback for a new essay by measuring the distance between features within the essay produced and a corpus of pre-scored essays, using an algorithm. If there is no record of a prompt in the corpus, the system will evaluate essays with a default scoring formula [6]: Total score = Vocabulary (43%) + Sentence (28%) + Structure (22%) + Content relevance (7%). However, there is no information about the rater identity and the scoring algorithm. After submitting an essay, *Pigai* can generate feedback containing three parts: (1) a holistic score; (2) general comments in terms of vocabulary, sentence, structure, and content relevance, and uses a bar graph to show the strength of the essay; (3) an analysis of linguistic features at the sentence level including errors, warnings, learning tips, and suggested usage. Despite its widespread use, there is little research on the validity evidence for improving writing ability. Therefore, this study aims to addresses this gap by adopting the framework for evaluation and use of AWE.

2 A Working Framework

A working framework of interpretive argument (Figure 1) was developed to evaluate validity of using *Pigai* in the EFL writing assessment.



Inference: Warrant *Assumptions (numbered under each warrant)*

E. Utilization: *Pigai's* feedback helps to promote students' writing ability.

1. Students hold positive views of using *Pigai* in writing assessment.
2. Students are willing to use *Pigai* in their writing process.
3. Teachers consider *Pigai* useful to promote students' writing ability.
4. Teachers are willing to let students use *Pigai* in their writing process.

D. Extrapolation: *Pigai* score is consistent with independent measures of students' writing ability.

1. *Pigai* score is consistent with teachers' rankings of students' writing ability.

C. Explanation: *Pigai* score reflects curriculum requirements of writing ability.

1. *Pigai* feedback contains rubric-relevant text features.
2. *Pigai* feedback does not contain rubric-irrelevant text features.

B. Generalization: *Pigai* score is consistent across various instances of the same error.

1. There is no significant difference between *Pigai* scores given to the same students with different tasks.

A. Evaluation: *Pigai* is accurate to score students writing performance.

1. *Pigai* score distributes consistently with the criterion score.
2. *Pigai* score correlates significantly with the criterion score.
3. *Pigai's* severity, internal consistency and bias against students are acceptable.

Fig. 1. Inferences, warrants, and assumptions in the validity argument for using *Pigai* in writing assessment

Three studies are conducted to collect evidence for assumptions in the validity argument. Two issues are worth mentioning here. First, as CET4 is the largest language test in China [7], the CET4 writing rubric was used because of its familiarity with students. Second, as *Pigai* doesn't reveal its scoring engine as *E-rater* does, we have to infer text features adopted by the system by analyzing its feedback.

3 Study 1

3.1 Research Purpose

This study aims to collect evidence of evaluation and explanation in the interpretive argument. Two research questions are raised: (1) What is the reliability of *Pigai* scores? (including A1, A2 and A3 in Figure 3) (2) Does the *Pigai* feedback include text features described in the CET4 rubric? (including C1 and C2)

3.2 Method

Materials and Instruments

CET4 writing adopts a holistic 15-point rubric including five score bands [7]. It describes four constructs including coherence, topic relevance, comprehensibility, and accuracy. It uses five scores (2-, 5-, 8-, 11-, and 14-point) to anchor raters' mental representation. In practice, the range finders (i.e., five benchmark essays provided by National College English Testing Committee (NCETC) to anchor raters' judgment) would be provided to guide rating training. 70 range finders between 2007 and 2014 were used because they were calibrated with preset scores by NCETC. After inputting these essays to the system, *Pigai* scores and feedback can be obtained.

Data Analysis

First, the Multi-faceted Rasch Model (MFRM) analysis for the ratings was conducted in FACETS version No. 3.58 [8]. Since CET4 essay rating adopts a holistic scale, then a two-facet mathematic model was built, where candidates and raters (including *Pigai* and the Criterion) were specified as facets.

$$Log(P_{ijk} / P_{ijk-1}) = B_i - C_j - F_k \qquad (1)$$

where $P_{ijk}$is the probability of examinee (i) being awarded a rating of (k) when rated by rater (j); P_{ijk-1} is the probability of examinee (i) being awarded a rating of (k-1) when rated by rater (j); B_i represents the ability of examinee (i); $C_j$represents the severity of rater (j); and F_k represents the step difficulty of being awarded a rating of (k) relative to (k-1) along the rating scale.

Second, feedback generated by *Pigai* was first segmented into independent "idea units" [9], then coded following guidelines of Grounded Theory [10]. In total, the feedback was segmented into 347 idea units. As a reliability check, all data were coded by a research assistant and the author separately. The inter-coder reliability reached 95.10%. Disagreements were resolved through negotiation.

3.3 Results

Evidence of Score Evaluation

Table 1 shows the descriptive results of *Pigai* and Criterion scores.

Table 1. Descriptive results of *Pigai* scores

	M	SD	Frequency				
			2-point band	5-point band	8-point band	11-point band	14-point band
Pigai score	7.86	3.16	3	20	28	4	15
Criterion score	8.00	4.27	14	14	14	14	14

As shown in Table 1, *Pigai* scores distribute more concentrated than Criterion scores, particularly in the 5-point and 8-point bands, suggesting the existence of central tendency. The inter-rater reliability was measured by the Spearman rank correlation coefficient as scores are not normally distributed. The result (ρ_s=0.865，n=70，*p*<.001) suggests that *Pigai* score has a fairly high inter-rater reliability.

MFRM results show that first, *Pigai's* severity (0.02 logits) is near to the Criterion score (-0.02 logits). Both rater separation ratio and reliability of rater separation index reached 0.00. The chi square test value (χ^2=.2, df=1, *p*>.05) also shows no significant difference between two groups in terms of severity. Second, the infit value of the *Pigai* score and the Criterion score reached 0.91 and 0.92 respectively, showing that *Pigai* has a good level of internal consistency. Last, the bias analysis revealed that *Pigai* has two biases towards essays, accounting for 1.43% of the total interactions (140) between raters and essays, which is acceptable [11].

Evidence of Score Explanation

Table 2 shows the frequency of each code in *Pigai's* feedback.

Table 2. Frequency of codes in the *Pigai's* feedback

Main category	Category	2-point band	5-point band	8-point band	11-point band	14-point band	Total
General impression	Fluidity	1	3	3	1	1	9
Structure	General evaluation	2	2	3	1	0	8
	Paragraphing reasonableness	5	2	2	5	0	14
	Convergence	4	1	0	1	1	7

	Compactedness	5	4	4	2	9	24
Language	Variety	17	16	17	15	9	74
	Complexity	10	20	18	24	26	98
	Appropriateness	1	1	3	0	1	6
	Accuracy	19	16	5	9	5	54
	Cohesion	15	11	8	9	10	53

According to Table 2, most codes are related with language, such as accuracy (19), variety (17), and cohesion (15), and no code with content, suggesting that *Pigai* focuses on language form. The four text features including coherence, topic relevance, comprehensibility, and accuracy in CET4 writing rubric are the de-facto intended constructs. It was found that rubric-related features only accounts for 33.43% in *Pigai's* feedback, implying the existence of construct irrelevance in *Pigai's* scores. Nonetheless, *Pigai's* feedback covers three kinds of rubric-related features, showing that *Pigai* scoring features represent the construct of interest to some extent. As *Pigai* adopts a heavy percentage of form-related features, issues like whether students would adopt certain form-dominated writing strategy are worth analyzing.

4 Study 2

4.1 Research Purpose

This study aims to collect evidence of generalization and utilization. Two research questions are raised: (1) Can *Pigai* score be generated to different tasks? (including B1) (2) What are students' attitudes towards *Pigai's* feedback? (including E1 and E2)

4.2 Method

Participants

Sixty-one EFL learners, and one EFL teacher participated in the study. These students, aging between 17 and 19, were from two intact classes (Class A and Class B) at a university in China. They were first-year undergraduate students and were enrolled in a freshman College English course. Class A has 16 males and 15 females, and Class B 14 males and 16 females. They were taught by one EFL teacher with over ten years' experience in teaching English as a foreign language.

Materials

Two writing tasks (Appendix 1 and 2) were selected as after-class assignment. They are typical CET4 writing tasks, which require students to write an argumentative essay with no less than 150 words.

Procedure

The study adopted a counter-balanced design across two weeks to control the order effect. Class A finished Task 1 in Week 1 and Task 2 in Week 2, while the order of tasks was reversed for Class B. At the end of research, a semi-structured interview was arranged on an individual basis. Two students (one male and one female) from each class were purposively chosen because of their willingness to participate. S1 to S4 were used to preserve their anonymity. The following questions were designed to guide students: (1) What effect does *Pigai* feedback have on your writing? (2) Are you willing to receive *Pigai* feedback in the future? Why?

Data Analysis

As the interval between two tasks is just one week, it could be operationally argued that students' writing ability does not change. Therefore, quantitative analysis was first conducted to determine descriptive statistics of *Pigai* scores for the two tasks, and the correlation coefficient between them. Then, a paired-sample t-test was conducted to test whether the means of *Pigai* scores for two tasks are equal. Finally, the interview protocols were analyzed thematically [12].

4.3 Results

Evidence of Score Generalization

Results of descriptive analysis suggest that *Pigai* scores of Task 1 (M=12.61, SD=0.56) is close to Task 2 (M=12.52, SD=0.65). The Kolmogorov-Smirnov test result (p>.05) shows that the two sets of scores are normally distributed. Pearson correlation coefficient between the two sets of scores reached .443 (n=61, p<.001). Results of the paired-sample t-test (t=1.00, df=60, p>.05) suggest that there is no significant difference between the two tasks.

Evidence of Feedback Utilization (Students' Attitudes)

Pigai's feedback was deemed useful particularly in the three aspects. First, it helps students identify errors quickly. For example, the following quote from S1 suggests that students appreciate *Pigai's* ability to detect errors. "*Pigai's feedback really helps, because I can quickly know errors in my essay, such as spelling error, phrase error and grammatical error.*" (S1). Second, *Pigai* provides information of highly scored essays, which improve students awareness of how to score high. In one example, S2 shared such experience. "*Pigai feedback is illuminating because it indicates clearly which expression is key to scoring high. I remember that Pigai pointed out that 'have ... confidence' is a common collocation and appeared 7,096 times in the corpus'. Since then, I began to use the structure.*" (S2) Last, *Pigai's* provision of referenced synonyms is deemed beneficial to enlarge students' vocabulary size. S4's comment below shows clearly that this information helps students vary their expressions. "*In Task 1, Pigai told me that 'prepare for' can be replaced by 'brace for' in that context, which was useful to improve the lexical capacity.*" (S4)

Nonetheless, the interviewees also queried the effectiveness of *Pigai's* feedback. First, *Pigai's* benefit is quite limited as it is not able to provide any information about content. A common disadvantage is pointed out by S1. "*Pigai's feedback focuses on language form such as spelling and collocation, while ignores other writing components like content, layout, and logic.*" (S1) Second, the most common negative perception is being too general to act upon. S4 commented clearly below. "*Pigai commented that my article does not read fluidly and advised me to use more linking words. However, it didn't specify the position. I was left puzzled.* " (S4)

Moreover, all students expressed their willingness to receive *Pigai's* feedback in the future because "*It can enhance my collocation ability*" (S2), "*I know some techniques how to achieve a good score*" (S3), "*I cannot receive such abundant and timely feedback from my teacher.*" (S4), and "*I can promote my lexical ability*" (S1).

5 Study 3

5.1 Research Purpose

This study aims to collect evidence of extrapolation and utilization. Two questions are raised: (1) What is the correlation coefficient between *Pigai's* scores and teachers' rankings of student writing ability? (including D1) (2) What are teachers' attitudes towards *Pigai's* feedback? (including E3 and E4)

5.2 Method

Participants
722 EFL learners and their seven EFL teachers (T1 to T7) participated in the study. These students, ranging in age from 17 to 19, were from 14 intact classes at a university in China. They were first-year undergraduate students and were enrolled in the same course like Study 2. Each teacher taught two classes. After writing on *Pigai* for one year, students and teachers were well informed of the *Pigai's* feedback.

Materials and Instruments
First, students' writing texts in the course exam of the first year were obtained. The task prompt can be found in Appendix 3. Second, teachers' rankings of these students' writing ability were solicited. Last, a questionnaire (Appendix 4) was administered to the seven teachers.

Data Analysis
First, the Spearman rank correlation coefficient was calculated to determine the relationship between *Pigai's* scores and teachers' rankings. Second, the quantitative part of teachers' response to the questionnaire (mainly Question 1 and 3) was analyzed descriptively. Last, the qualitative part of teachers' response to the questionnaire (mainly Question 2) was analyzed following Grounded Theory [10]. As a result, teachers' responses can be segmented into 62 idea units. Those codings for and against using *Pigai's* feedback amount to 33 and 29 respectively. The coding reliability between the research assistant and the author reached 93.55%, suggesting the creditability of coding results.

5.3 Results

Evidence of Score Extrapolation
It was found that the Spearman rank correlation coefficients between *Pigai's* scores and teachers' rankings for each class ranged between 0.39 (p<.01) and 0.70 (p<.01), which suggested that *Pigai* scores have substantial relationship with teachers' rankings. This result was also cross-validated by teachers' responses to Question 1 in

the questionnaire, where all teachers considered that *Pigai* score was largely consistent with their observations of students' writing ability.

Evidence of Feedback Utilization (Teachers' Attitudes)

Teachers' attitudes toward using *Pigai* in the classroom can be summarized with Table 3. All teachers expressed their willingness to let students receive *Pigai's* feedback as far as Question 3 in the questionnaire is concerned.

Table 3. The coding framework of teachers' attitudes

Code	Example
Advantage	
1. Able to identify errors	"*Pigai* is able to diagnose some spelling errors." (T1)
2. Enrich assessment methods	"As teachers can set sample essays for students' reference, and students can realize their disadvantages during the revising process, *Pigai* thus enriches the assessment method." (T5)
3. Improve vocabulary	"*Pigai'*s feedback on synonyms is quite useful." (T3)
4. Develop the habit of revising	"Writing on *Pigai* can help students revise their own performances from time to time, which is good to form a good habit." (T7)
5. Stimulate interest	"Students become more interested in writing." (T4)
6. Facilitate learning by revising	"A student revised the essay over 70 times on *Pigai*. He knows his writing problems more deeply in the process, which cannot be achieved by relying on teachers' corrective feedback." (T7)
7. Enable to write more	"Students have more chances to write, as compared with the traditional writing instruction." (T2)
Disadvantage	
1. Difficult to understand	"*Pigai's* feedback on grammar is sometimes puzzling." (T6)
2. Difficult to act upon	"Students have no idea how to revise based on *Pigai's* feedback because some of them are too general and ambiguous." (T2)
3. Inaccurate judgment	"Some grammar feedback contains erroneous information." (T6)
4. Narrow down the construct	"*Pigai* feedbacks only on form-related features, which are just part of writing ability." (T2)
5. Limited to boost learning	"Relying solely on *Pigai's* feedback can have a limited effect to promote students' writing." (T4)

6 Discussion and Conclusion

The main findings of the research are summarized below:

First, *Pigai* yields scores that are accurate indicators of the quality of a test performance sample (including Assumptions A1, A2, and A3), *Pigai* yields scores that are sufficiently consistent across tasks in the same form (including B1), and *Pigai* yields scores that are consistent with teachers' judgments of students' writing ability (including D1). However, *Pigai* scores tend to be more centralized and distribute more narrowly than the criterion scores. There are some possible reasons. First, *Pigai's* scoring features are predictive of scores awarded by human raters. As *Pigai* derived the score of an essay based on a large corpus of human-scored essays, the scoring algorithm can help *Pigai* extract distinctive features and ensure its reliability. Second, the task prompts used in this study are with similar genre and structure, which helps *Pigai* achieve a good reliability across prompts. Finally, as CET4 writing rubric emphasizes language rather than content, students would give priority to producing texts with accurate language. Under the context where all AWE system can only judge on surface features, *Pigai's* scoring reliability would be improved.

Second, *Pigai* scoring features represent the construct of interest to some extent, yet problems of construct under-representation and construct-irrelevance still exist. As *Pigai's* feedback is deemed general and opaque by most users, its effect on improving students' writing ability is doubtful. It would be better for *Pigai* to develop both general and prompt-specific modeling for scoring. In addition, *Pigai* is suggested to provide clear definition and specific example of certain text features in the feedback, such as "convergence" and "compactedness" in Table 2.

Finally, *Pigai* generates feedback that helps students' development of writing ability, but to some extent (including E1 and E4). The root cause may lie in the feedback explanation of *Pigai.* Fundamentally, a computer cannot score essays in the same way as a human rater. It generates scores by devising certain algorithm using natural language processing and so on, rather than drawing on certain learning theory or writing theory. Therefore, there are still a number of doubts and oppositions against its application to L2 writing assessment [1]. As conceptualization of the writing construct is narrowed down using an AWE system, students may develop a primarily formalist approach to writing, i.e. writing to a machine rather than writing to a human. In that case, the authenticity of writing instruction and assessment would be seriously violated. Considering that AWE can never replace the role of human in the writing assessment, students should be trained to conduct other forms of assessment such as peer assessment and self-assessment for their writing.

There are several limitations. First, all inferences focus on scores except utilization, which are concerned with feedback. Therefore, investigating *Pigai's* feedback in terms of evaluation, generalization, explanation and extrapolation is warranted. Second, as all the task prompts are with the same genre (i.e. argumentative), the study should be replicated with different text types. Last, none of the sub-studies provided the result related to the system's effectiveness on affecting students' writing performance, which should be investigated further in the future.

References

1. Warschauer, M.: Automated writing evaluation: Defining the classroom research agenda. Language Teaching Research **10**, 1-24 (2006).
2. Valenti, S., Neri, F., Cucchiarelli, A.: An overview of current research on automated essay grading. J. Inf. Technol. Educ. Res. **2**, 319-330 (2003).
3. Xi, X.: Automated scoring and feedback systems: Where are we and where are we heading? Language Testing **27**, 291-300 (2010).
4. Williamson, D. M., Xi, X., Breyer, F. J.: A framework for evaluation and use of automated scoring. Educational Measurement: Issues and Practice **31**, 2-13 (2012).
5. Zhang, Z.: Student engagement with computer-generated feedback: A case study. ELT Journal **70**, 1-12 (2016)
6. Bai, L., Hu, G.: In the face of fallible AWE feedback: How do students respond? Educational Psychology **37**, 67-81 (2017)
7. Zhang, J.: Same text different processing? Exploring how raters' cognitive and meta-cognitive strategies influence rating accuracy in essay scoring. Assessing Writing **27**, 37-53 (2016).
8. Linacre, J. M.: A user's guide to FACETS: Rasch-model computer programs. MESA Press, Chicago (2005).
9. Green, A.: Verbal protocol analysis in language testing research: A handbook. Cambridge University Press, Cambridge (1998).
10. Glaser, B. G., Strauss, A. L.: The discovery of grounded theory: Strategies for qualitative research. Aldine de Gruyter, Chicago (1967).
11. McNamara, T. F.: Measuring Second Language Performance. Longman, London (1996).
12. Miles, M. B., Huberman, A. M.: Qualitative data analysis: An expanded sourcebook. Sage, Thousand Oaks, CA (1994).

Appendix 1: Task 1 (A Technological Invention)

Write an essay of no less than 150 words about a technological invention. Your writing should include four points: 1. An introduction of the invention. 2. Its positive impact on peoples' life. 3. Its negative impact on people's life. 4. Your opinion.

Appendix 2: Task 2 (Fame – Good or Evil?)

Write an essay of no less than 150 words on the topic "Fame-Good or Evil?" Your paper should cover the following points: 1. The advantages of being famous. 2. The disadvantages of being famous. 3. Your attitude towards fame.

Appendix 3: The Internet and Our Daily Lives

Write an essay of no less than 150 words on the topic "The Internet and Our Daily Lives". Your paper should include: 1. Internet is important in our daily lives. 2. Internet has also disadvantages. 3. What shall we do to make better use of Internet?

Appendix 4: A Questionnaire of Teachers' attitudes towards *Pigai*

1. Is *Pigai* score consistent with your observation of students' writing ability?

A. Consistent B. Largely consistent C. Largely inconsistent D. Inconsistent

2. Is *Pigai's* feedback beneficial to improve students' writing ability? Why?

3. Would you like to let students receive *Pigai's* feedback in the future? (Yes/No)