Exploring the Feedback Quality of an Automated Writing Evaluation System *Pigai*

https://doi.org/10.3991/ijet.v16i11.19657

Jianmin Gao
Zhejiang University, Hangzhou, China
jimmy_gao@zju.edu.cn

Abstract—The study investigated the feedback quality of an Automated Writing Evaluation system (AWE) *Pigai*, which has been widely applied in English teaching and learning in China. The study not only focused on the diagnostic precision of the feedback but also investigated the students' perceptions of the feedback use in their daily writing practices. Taking 104 university students' final exam essays as the research materials, the paired sample t-test were conducted to compare the mean number of errors identified by *Pigai* and professional teachers. It was found that *Pigai* feedback could not so well diagnose the essays as the human feedback given by the experienced teachers, however, it was quite competent in identifying lexical errors. The analysis of students' perceptions indicated that most students thought *Pigai* feedback was multifunctional, but it was inadequate in identifying the collocation errors and giving suggestions about syntactic use. The implications and limitations of the study were discussed at the end of the paper.

Keywords—Feedback quality; Automated Writing Evaluation system; *Pigai*

1 Introduction

With the development of computer and information science, the automated writing evaluation systems (AWE) have been drawing more and more attention from researchers in language teaching and assessment. Since it has many advantages over human assessment in its high efficiency, high consistency, and low cost [1-2], many high-stakes tests have included AWE in their rating process. For example, both the Graduate Record Examination (GRE) test and the Test of English as a Foreign Language (TOEFL) have adopted a mixed use of AWE and human scoring. The prevailing use of AWE in the high-stakes tests of English proficiency also induced a series of investigations into the reliability and validity of the score offered by an AWE, and many research indicated that the combination of the AWE scoring and human scoring in the assessment of writing performance could reflect test-takers' language proficiency in a precise way [3-5]. In fact, apart from simply giving a holistic score of an essay based on the underlying algorithm, many AWE systems can also provide users with a systematic feedback on the essay quality, which is not only quite helpful in assisting instructors to locate language weaknesses of language learners but also constructive in facilitating learner' self-diagnosis. Many efforts have been made in investigating the

effectiveness of AWE feedback, but they were mainly focused on the AWE developed in western countries, such as MY Access and Criterion, however, these systems are not easily accessible in China. Since China is a country with almost the largest population of English as a foreign language (EFL) leaners, the study investigating the feedback quality of the AWE that is widely applied in China is of vital importance. Among all the AWE available in China, *Pigai* is the first system used to facilitate English teaching and learning, and it has gained a lot of popularity among teachers and students in both middle schools and universities. Therefore, the current study is dedicated to evaluating the quality of the feedback provided by *Pigai*, aiming to offer suggestions for the users of this system and inspire more empirical research in validating the effectiveness of AWE widely applied in EFL countries.

1.1 An introduction of *Pigai*

Developed by a company in Beijing (China), *Pigai* is an online system that applies computer algorithm to automatically score English essays. Just like a doctor using a CT machine, language instructors can automatically scan the various parameters related to a student's essay and make more accurate and objective judgments and comments.

The working mechanisms of *Pigai* is to compare the difference between students' essays and the standard corpus, and then map such information into scores and comments by a certain algorithm. This technique in automatically scoring essays and giving feedbacks has been patented in China. The developers of *Pigai* hold the belief that the specific feedbacks and suggestions are more important than essay scores, for they enable the users to improve their essays in a correct way.

By using *Pigai*, students can first get a holistic score and a summative comment on their language performance. In addition, they can get a specific feedback in which every sentence with linguistic errors will be given a corrective suggestion. A learning tip related to the lexical and syntactic use will also be offered. As Table 1 shows, *Pigai* is a multifunctional tool that can not only identify different types of linguistic errors but also provide suggestions on the future learning.

Pigai has already become a teaching-assistant tool in China. According to the statistics presented in the official website of *Pigai*, approximately 400 million essays have been corrected and evaluated on this platform up to 2018.

Table 1. The major functions of Pigai automated writing evaluation system

Functions	Descriptions
Identifying spelling errors	Errors in spelling, capitalizing and the use of punctuations
Identifying content words-related errors	Errors in using the morphological form of nouns, verbs, and pronouns; errors in ranking the order of different adjectives and adverbs; misusing adjectives as adverbs and vice versa.
Identifying function words-related errors	Misusing or lacking articles, prepositions, and conjunctions.
Identifying collocation errors	Grammatically incorrect collocations (e.g., too much things); Non-native expressions
Identifying syntactic errors	Errors leading to an incomplete sentence structure
Giving tips on lexical use	Synonym analysis; collocation suggestions
Giving tips on syntactic use	Sentence structure suggestions

2 Literature Review

Most of the studies that investigated the effectiveness of AWE feedback attached great importance to the variation in the writing proficiency of students after they used the feedback as a learning tool. Researchers first applied a pre-test to record the writing proficiency of students before they used the AWE feedback and then a post-test after several weeks or months to observe the difference in the writing scores and textual features of the students' essays after they used the AWE feedback to facilitate their learning. Through such a comparison, researchers could make the inference about whether the AWE feedback can improve the writing quality in a comprehensive way or it only has a very limited effect. Liao [6] investigated the effect of AWE feedback on improving grammatical accuracy by observing the changes in the quality of students' essays over time, and the result showed that errors in broken sentences and subject-verb agreement decreased significantly, but the number of sticky sentences and verb-related morphological errors did not see a statistically significant variation, suggesting that the effect of online automatic feedback on language accuracy was constrained by the type of errors. In Lv's research [7], it was also found that the *Pigai* feedback was only limited in decreasing the grammatical errors and improving the overall score of students' essays, and it nearly has no effect on increasing writing fluency and improving essay structure. However, studies like these have an inevitable disadvantage in the methodological aspect. During the process, students' overall language proficiency also increases as they take lessons, do homework, or input more English materials, therefore, it is difficult to completely attribute the improvement of the students' essays to the use of feedback. Some research also compared the effect of AWE feedback and teacher feedback. For example, Dikli and Bleyle [8] found that AWE feedback did not identify the grammatical errors so well as the teacher feedback did.

Apart from the research exploring the effect of AWE feedback on improving writing quality, some researchers also investigated the perceptions of students about the use of feedback. Students have mixed opinions on whether the feedback is effective, and most studies showed that learners expressed disappointment and mistrust in feedback [9]. For example, Grami [10] suggested that students who were the frequent users of online corpora were more inclined to use AWE feedback, and regardless of the quality of the feedback, students preferred the tools that enable them to get an instant feedback. In Chen and Cheng's research [11], 50% of the respondents did not think that online automatic feedback was helpful in improving their writing. Students thought that online feedback was too simple and mechanical and only suitable for low-proficiency students, while the feedback on content quality and creativity were more important for high-proficiency students. Shermis et al. [12] also reported that only 112 students (21% of the total) submitted all seven writing tasks using the online system during a 20-week writing instruction period, suggesting a low motivation of students in using the AWE. In addition, more studies have found that students' enthusiasm for the use declines over time [13].

Above all, it could be noticed that most studies are focused on how much can AWE feedback help improve the writing quality and how AWE users feel about the

feedback. However, rarely seen is the research investigating the quality of feedback itself, that is, to what extent can the feedback accurately diagnose the students' essays. Besides, most research only concerned the students' overall impression about the use of feedback instead of asking students to report their feelings about each function claimed by AWE, which might lead to an incomplete report of the students' perceptions. Taking all these research gaps into account, the current study is dedicated to investigating the quality of the *Pigai* feedback and students' perceptions of each function of this AWE. The two specific questions are:

1. How does the *Pigai* feedback differ from the feedback given by the experienced teachers? Or to what extent can the *Pigai* feedback precisely and comprehensively reveal the essay problems?
2. How do students feel about the effectiveness of the *Pigai* feedback in terms of the each function claimed by this AWE?

3 Method

3.1 Materials and instruments

The research materials in this study were 104 students' essays. These essays were finished in the final exam of an English course mainly centered on training students' writing and speaking skills. All the students finished the exam in 100 minutes, and according to the instructor of this course, there was not an obvious difference in their writing speed based on the daily observation. Since the score of the exam was closely related to the final grade of the course, it could be inferred that all the students had been concentrated on their writing during the test. The topic of the independent writing task was related to the description and analysis of a social problem.

In this study, 5-point Likert scale were adopted to investigate the students' perceptions of using the AWE feedback in their weekly writing practice. A question concerning the use frequency was designed, and the data given by the students who chose the option "I never use it" were excluded. In addition to this question, the survey consisted of 7 items, with each item investigating students' perception of the extent to which each function listed in Table 1 was fulfilled.

3.2 Data collection

Two experienced teachers participated in the study. They had a random selection of 10 essays from the 104 in total to have the first pilot coding. All the errors in the essays should be found and classified into different types and then quantified into the error type frequency. The two teachers independently coded these essays and then made a careful comparison between the corrective feedback they had made. After a thorough discussion, they had the second pilot coding of 14 essays which were randomly selected from the rest 94 essays. After they reached a high consistency between their identified error types and the corresponding frequency (r = .82, p <.05), they

split the rest essays and each had an independent coding. Apart from the human feedback, the AWE feedbacks provided by *Pigai* were also collected.

All the students finished the survey online one week before they had their final exam. This time design was used to reduce the influence of their testing anxiety on their perceptions, and it could also alleviate the influence of their impressions on the course instructor, for some students who scored low in the final exam might be unwilling to finish the survey. After excluding the invalid data, 102 students' responses were collected.

3.3 Data analysis

For the first research question, to what extent can the *Pigai* feedback precisely and comprehensively diagnose the essay problems, paired sample t-test was used to compare the mean number of errors of different types identified by the *Pigai* and the the professional teachers comparison, the discrepancy in the quality of essay evaluation between two kinds of feedbacks could be recognized. In addition, with an reference to Leacock et al. [14], the researcher also calculated the precision rate and the recall rate of the *Pigai* feedback. The former indicator was used to see the error-picking precision, and the latter was used to judge the error-picking coverage. Precision rate was calculated by dividing the number of the errors that were accurately identified in the *Pigai* feedback by the total number of the errors identified by it. In order to calculate the precision rate, the two teachers together made a detailed check of the *Pigai* feedback for the unsuccessful corrections after they finished giving their own feedbacks. The Recall rate was calculated by dividing the number of the errors that were accurately identified in the *Pigai* feedback by the total number of the errors identified by the professional teachers.

For the second research question, a Bar chart was used to illustrate students' perceptions of the different functions of *The Pigai* feedback.

4 Results and Discussion

For the first research question, the descriptive data presented in Table 2 and the paired-samples t-test results suggested that in *Pigai* and human feedbacks, there was no difference in the number of the spelling errors (p = .165), content words-related errors (p = .165), and function words-related errors (p = .721). While they did share the significant difference in the number of collocation errors (p = .001), syntactic errors (p < .001) and logic problems (p < .001). The precision rate and recall rate of *Pigai* feedback were 94.42% and 71.58%.

Table 2. Descriptive statistics of the mean number of the errors in different types

Error types	*Pigai* feedback	Human feedback
Spelling errors	3.15(2.44)	3.31(2.40)
Content words-related errors	4.69(1.55)	5.00(1.63)
Function words-related errors	5.31(2.14)	5.23(2.13)
Collocation errors	2.69(0.95)	3.85(0.99)
Syntactic errors	2.31(1.18)	4.31(1.11)
Logic problems	0	1.85(1.28)

The results above indicated that although *Pigai* feedback could well identify the spelling errors and most lexical errors as the professional instructors, it was not quite helpful in picking out collocation and syntactic errors, which are more complex in that these types of errors concern the knowledge of combining single words into phrases and sentences. In addition, *Pigai* could not identify the logic problems at all, which were found to be common in the students' essays by both of the two teachers. This corroborated the research of Dikli and Bleyle [8] in that AWE feedback were merely effective in reflecting the errors related to the lower-order language skills while less helpful in revealing the deficiency in managing higher-order skills. Logic management is one of the higher-order skills, since it concerns how students link different sentences together to form a coherent and cohesive discourse. In addition, in Hoang and Kunnan's research [15] in evaluating MY Access feedback, it was also found that the AWE feedback was not sufficient in recognizing organizational problems at the discourse level and off-topic content. This suggested that AWE feedbacks were still limited in its diagnostic effectiveness.

According to Lv [7], interactive hypothesis theory plays an important role in the teaching of a foreign language. In fact, the theory can also provide us with a perspective to understand why AWE feedback could not diagnose the logic problems. Since the writing process is also a dynamic and interactive process, one task writers ought to finish is to interact with their target readers, in this case, being logical in their writing outcomes is of great importance. Once the essays failed to be clear or understandable, the target readers who were the instructors in this study could immediately identify the logical problems, however, AWE works on the mere basis of the computer algorithm, which is not able to interact with a human being.

For the second research question, Table 3 and Fig. 1 showed students' mixed opinions about the extent to which *Pigai* feedback had fulfilled its various functions. In order for a clear illustration, the 7 functions listed in Table 3 were named as F1 to F7 in Fig. 1. If we take 4-5 as the mid-interval in a 7-point scale, as was shown in Table 3, students held a positive attitude to its fulfillment of functions, namely, identifying spelling errors (6.41), giving tips on lexical use (6.35), identifying function words-related errors (6.06) and identifying content-related words (5.76), while they were negative about its function of giving tips on syntactic use. As was illustrated in Fig.1, students tended to acknowledge *Pigai*'s functions in correcting single words-related errors (F1, F2, F3, and F6), while they were unwilling to admit its effect in helping improve sentence-level competence (F4, F5, and F7). An influential research conducted by Warschauer and Grimes [16] also pointed out that students' self-corrections

of their writings were mainly at the word level, such as spelling, punctuations and grammar, while the revisions in content and discourse organization were rare. To some extent, this indicated the students' mistrust in AWE feedback's functions in identifying higher-order writing problems. The results illustrated in Table 3 and Fig. 1 were also in accordance to the results of the first research question. There were two possible explanations. One is that *Pigai* was indeed less successful in identifying the sentence-level errors and offering suggestions on syntactic use. Another reason was that the students involved in this study were limited in their English writing proficiency that they were not able to self-reflect their deficiency according to the *Pigai* feedback, and this needs more deeper post-interviews in future studies.

Table 3. Descriptive statistics of the function fulfillment scores

Functions	Min	Max	Mean	SD
F1: Identifying spelling errors	5.00	7.00	6.41	0.80
F2: Identifying content words-related errors	4.00	7.00	5.76	0.97
F3: Identifying function words-related errors	5.00	7.00	6.06	0.75
F4: Identifying collocation errors	3.00	7.00	4.24	1.09
F5: Identifying syntactic errors	2.00	6.00	4.35	1.06
F6: Giving tips on lexical use	5.00	7.00	6.35	0.70
F7: Giving tips on syntactic use	1.00	6.00	3.53	1.42

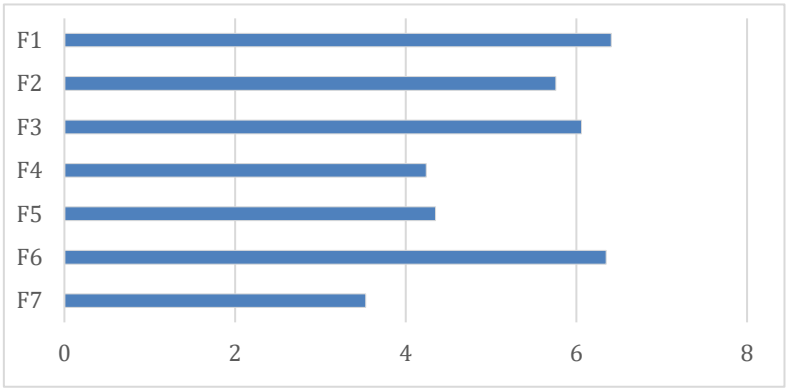

Fig. 1. The distribution of the function fulfillment scores

5 Conclusion

In this study, the quality of an AWE feedback was examined. To be specific, the extent to which the feedback can precisely and comprehensively identify the language errors and the extent to which the feedback had fulfilled each of its 7 functions perceived by its users were examined. The results showed that although the feedback provided by *Pigai* was quite precise in its judgment of errors, it could not well identify the language errors in all aspects, and it was merely able to well diagnose the lexical errors and give suggestions on lexical improvement. These results were consistent

with the students' opinions about its functions in that they were more positive about the functions in improving words-related aspects while relatively more negative about the functions in promoting syntactic improvement. These results suggested that *Pigai* or other AWE feedbacks should be used with caution, although they have the advantages of high efficiency and consistency, it should be applied with the combination of other feedbacks.

The study also has some limitations that have to be admitted. First, the study merely involved two teachers to give essay feedbacks, although they were both trained and experienced, producing a standard feedback for making comparison with AWE feedbacks might take a larger amount of human coders. Second, the essays were all argumentative writings in this study. Therefore, the generalization of the findings in this study to other types of writings should be treated with caution. Finally, a more detailed post-interview about students' perceptions of the use of AWE feedbacks should be designed in the future study.

6 References

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

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

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

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

[5] 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

[6] Liao, H. (2016). Using automated writing evaluation to reduce grammar errors in writing. ELT Journal, (3): 308-319. https://doi.org/10.1093/elt/ccv058

[7] Lv, X. (2018). A Study on the Application of Automatic Scoring and Feedback System in College English Writing. International Journal of Emerging Technologies in Learning, 13(3): 188-196. https://doi.org/10.3991/ijet.v13i03.8386

[8] Dikli, S., Bleyle, S. (2014). Automated essay scoring feedback for second language writers: How does it compare to instructor feedback. Assessing Writing, (22): 1-17. https://doi.org/10.1016/j.asw.2014.03.006

[9] He, H. A Survey of EFL College Learners' Perceptions of an On-Line Writing Program. (2016), International Journal of Emerging Technologies in Learning, 11(4): 11-15. https://doi.org/10.3991/ijet.v11i04.5459

[10] Grami, M. A. J. (2020). An Evaluation of Online and Automated English Writing Assistants: Collocations and Idioms Checkers. International Journal of Emerging Technologies in Learning, 15(4): 218-226. https://doi.org/10.3991/ijet.v15i04.11782

[11] Chen, C., Cheng, W. (2008). Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes. Language Learning & Technology, (2): 94-112.
[12] Shermis, M.D., Burstein, J., Bliss. L. (2004). The impact of automated essay scoring on high stakes writing assessments. The annual meeting of the National Council on Measurement in Education, San Diego, CA.
[13] Scharber, C., Dexter, S., Riedel, E. (2008). Students' experiences with an automated essay scorer. The Journal of Technology, Learning and Assessment, 7(1). https://files.eric.ed.gov/fulltext/EJ838623
[14] Leacock, C., Chodorow, M., Gamon, M., Tetreault, J. (2010). Automated grammatical error detection for language learners. Synthesis Lectures on Human Language Technologies, 3(1): 1-134. https://doi.org/10.2200/s00275ed1v01y201006hlt009
[15] Hoang, G., Kunnan, A. (2016). Automated Essay Evaluation for English Language Learners: A Case Study of *MY Access*. Language Assessment Quarterly, 13(4): 359-376. https://doi.org/10.1080/15434303.2016.1230121
[16] Warschauer, M., Grimes, D. (2008). Automated writing assessment in the classroom. Pedagogies: An International Journal, 3: 22-36. https://doi.org/10.1080/15544800701771580

7 Author

Jianmin Gao is studying in the department of linguistics at Zhejiang University (310058), which is situated in Hangzhou, China. The author's research interest lies in second language acquisition and language testing.

Article submitted 2020-11-04. Resubmitted 2021-02-17. Final acceptance 2021-02-17. Final version published as submitted by the authors.