Sino-US English Teaching, October 2021, Vol. 18, No. 10, 267-273
doi:10.17265//1539-8072/2021.10.001

A Comparative and Integrated Study of English Composition Online Automatic Scoring (OAS) and Teacher Scoring (TS)

LI Wanjun, ZHAO Yun, JIA Wenfeng, ZHAO Yushan
Shandong University, Weihai, China

The objective of this paper is to explore the reliability of Online Automatic Scoring (OAS) through the comparison of OAS and Teacher Scoring (TS), and further demonstrate the feasibility of the integration of the two scoring methods. The Pearson correlation statistics of the two scoring results of 115 compositions are processed with SPSS analysis software, indicating that the correlation between the two reaches 0.83, which means that OAS is relatively reliable in dealing with students' compositions. After the second stage of the TS experiment, the questionnaire results show that students generally recognize the OAS and have a clear understanding of the advantages and disadvantages of the two scoring methods. Combined with the students' interview, the conclusion is that the OAS is reliable and the integration of the two scoring methods will have a better effect.

Keywords: online automatic scoring, teacher scoring, integration

Introduction

At present, many colleges and universities in China have adopted the OAS system to evaluate and offer feedback on students' English composition, which undoubtedly greatly alleviates the pressure of composition correction of English teachers, so that they can spend more time and energy on other aspects of teaching. The OAS system of English composition has its own advantages, such as high efficiency, fast information feedback, and strong objectivity (not affected by emotion). But its weakness is also obvious, that is, the computer cannot judge an article or appreciate an article like people, but can only do the corresponding work according to the program set by people. Therefore, the extent to which the OAS results can truly reflect students' composition level and put forward specific revising suggestions to students remains to be further verified. The objective of this study is to compare the Online Automatic Scoring (OAS) system of English composition with the traditional Teacher Scoring (TS), explore the reliability of the OAS system of English composition, and how to effectively integrate the two methods, so as to achieve a good combination of traditional teaching methods and modern educational technology, promote the further development of writing teaching, and effectively improve students' English writing.

Funding information: Funded by Shandong Social Science Planning Program (山东省社科规划项目).
LI Wanjun, Ph.D. candidate, associate professor, School of Translation Studies, Shandong University, Weihai, China.
ZHAO Yun, MA, Ph.D. candidate, lecturer, School of Translation Studies, Shandong University, Weihai, China.
JIA Wenfeng, Ph.D., lecturer, School of Translation Studies, Shandong University, Weihai, China.
ZHAO Yushan, MA, Ph.D. candidate, lecturer, School of Translation Studies, Shandong University, Weihai, China.

Literature Review

Related Researches Abroad

The OAS system was first developed by Ellis Page of Duke University in 1966, named PEG (Project Essay Grade). The system only evaluates the language quality in the composition, and only analyzes the surface features of the text, but does not evaluate the composition content (Page, 2003). For various reasons, the system did not make much progress in the next 30 years. In 1997, the University of Colorado in the United States developed an automatic composition scoring system IEA (Intelligent Essay Assessor), which uses the semantic text analysis in latent semantic analysis proposed by psychologist Thomas Landauer to evaluate compositions (Landauer, Laham, & Foltz, 2000). Its advantage lies in being able to evaluate the quality of text content, but its disadvantage is that it cannot analyze the language quality and text structure of the composition. Then, the Educational Testing Service (ETS) of the United States developed the e-rater composition scoring system based on natural language processing technology, information retrieval technology, and statistical technology, and then applied it to the composition marking of GMAT and TOEFL. The advantage of the system is that it focuses on analyzing the language, content, and text structure of the text. These three modules are more consistent with the teacher scoring elements. Its deficiency lies in the weak analysis of the content quality of the composition, the analysis of the text structure is confined to the surface characteristics of the text, and the analysis of the language quality is not comprehensive enough (Liang & Wen, 2007). In addition, the biggest problem of the system is that it cannot distinguish compositions with correct grammar but empty content (Chen & Ge, 2008).

Related Researches at Home

In China, Liang Maocheng (2005) of China Foreign Languages Research Center was the first to set foot in the research of automatic English composition scoring system. His doctoral dissertation studies the construction of automatic scoring model for Chinese students' English composition. His modeling method takes into account the advantages of PEG and IEA. His research has achieved high scoring accuracy, and the correlation coefficient *r* with teacher scoring is up to 0.873. However, due to the narrow source range and small number of composition samples, and the extracted features being mainly shallow features of the text, which cannot involve the deep structure of the article, the results need to be further verified and strengthened (Chen & Ge, 2008).

Ge and Chen (2007a; 2007b; 2009) were another group of early researchers in China who studied automatic English composition scoring. In addition to introducing foreign automatic composition scoring systems, they also conducted follow-up research and reported on domestic English composition scoring systems, and put forward relevant problems related to automatic scoring of compositions of domestic college English learners, for example, the pertinence of automatic composition scoring, the universality of automatic composition scoring, the division of man-machine interface of automatic composition scoring, etc. In addition, Wu and Zhang (2011) of Beijing University of Technology have also conducted a comparative experiment between machine scoring and teacher scoring, but whether the experimental results are repeatable remains to be tested. In addition, their experiment also involves the objectivity of teacher scoring, as well as the index and standard selected to prove the correlation, which are worthy of further discussion.

In addition, Zhou, Fan, Ren, and Yang (2021) discussed how to improve the effect of online composition scoring by obtaining multi-level semantic features such as deep semantic features and shallow linguistic

features through natural language processing technology. While Gao (2021) found that the consistency between the score of the online composition scoring system and that given by experienced teachers is low, it cannot fully reflect the language characteristics of the text.

Research Design

Research Questions

1. Does high correlation or consistency exist between the OAS system and the TS results through a class of students' composition practice experiment?

2. If the correlation is proved to be high, can the two scoring methods be effectively combined in order to give full play to the best combination of traditional teaching methods and modern educational technology?

Research Subject

The participants in this experiment are 39 first-year non-English majors in a class in a university in East China. They completed three compositions and submitted them to the OAS system in the first semester. At the same time, three teachers were arranged to score the compositions submitted by the students independently. The OAS system received a total of 115 valid compositions (two students did not submit composition once). These students also had to complete the practice of three compositions in the second semester, but they would have the TS only, not just scoring, but also normal teacher feedback (usually pointing out the strengths, weaknesses, and suggestions for improvement).

Research Methods and Procedures

The research team selects a complete class of 39 first-year students as the experimental class, and arranges three compositions in one semester, which will be scored by the OAS and TS respectively. In order to ensure the reliability and validity of the TS, three teachers are invited to score according to the scoring standard of College English Test band 4 (CET-4) and College English Test band 6 (CET-6) for compositions (full score being 15 points), and then the scores of the three teachers will be averaged to obtain a relatively objective score. With the scores of both OAS and TS, the correlation between them will be obtained through SPSS statistic software. If the correlation between the two is not high, it shows that the OAS system has defects in reliability and cannot be widely used. If the correlation between the two is proved to be high, the second stage experiment will be carried out, that is, TS will be completely adopted in the second semester, and then questionnaire survey and interview will be conducted on the experimental subjects to see whether the OAS system can completely replace the TS. If not, it will further explore how to effectively integrate the two scoring methods to make up for their shortcomings, so as to not only give full play to the advantages of the OAS system, but also reduce the heavy work of teachers.

Research Results and Analysis

Comparative Analysis of the OAS and TS Results

The research team collected totally115 compositions from the 39 first-year non-English majors, and all the compositions were scored by the OAS system and TS respectively in the way of holistic scoring. In order to obtain the correlation coefficient between the OAS and TS, the research team compared the scores of 115 compositions, and analyzed the results with SPSS statistics software, which is shown in Table 1.

Table 1

Pearson Correlation Results

	Average score	Standard deviation	OAS	TS
OAS	11.124	1.422	1	
TS	10.382	1.374	0.830**	1

* $p < 0.05$ ** $p < 0.01$

It can be seen from the above table that the correlation coefficient between OAS and TS is 0.83, with the significance of 0.01 level, which shows that there is a significant positive correlation between the OAS and TS.

Questionnaire Survey of OAS and TS

After one semester of OAS, the class adopted the TS method in the second semester. Teachers generally give scores and error feedback. At the end of the semester, a questionnaire was given to 39 students in the class, and 38 valid answers were obtained (one did not participate). In addition to the questionnaire, six of the 39 students were interviewed at random. The results are as follows:

The questionnaire of OAS and TS.

1. What do you think of the comparison between OAS and TS?

A. OAS is better than TS

B. TS is better than OAS

C. It is hard to say which is better since each has its own characteristics

2. What do you think is the biggest advantage of OAS?

A. Fast scoring and high efficiency

B. Scoring objectively and accurately

C. The error correction hints are accurate and rich

3. What do you think is the biggest deficiency of OAS?

A. The comments are too general and not targeted

B. The score was not objective

C. The error correction hints are unreasonable

4. What do you think is the biggest advantage of TS?

A. The comments are appropriate and targeted

B. Scoring objectively and accurately

C. The comments are encouraging

5. What do you think is the biggest disadvantage of TS?

A. Correcting the composition is too time-consuming, which inhibits the teacher's enthusiasm in arranging the composition

B. Scoring is not objective enough and may be affected by teachers' emotions and other factors

C. If the feedback of error information is too slow, it will affect students' enthusiasm for revision

6. Do you think TS can be effectively combined with OAS?

A. Yes B. Not sure C. No

Questionnaire results and analysis.

Table 2

Results of the Questionnaire Survey

Question choice	A (%)	B (%)	C (%)
1	15.8	18.4	65.8
2	73.7	15.8	10.5
3	78.9	7.9	13.2
4	42.1	31.6	26.3
5	39.5	21	39.5
6	84.2	7.9	7.9

As can be seen from the results of the questionnaire, most students believe that OAS can be combined with TS (65.8%). The vast majority of students (73.7%) believe that the biggest advantage of OAS is its fast scoring speed and high efficiency. For the biggest deficiency of OAS, 78.9% of the students thought that the comments were too general and not targeted. For the greatest advantage of TS, 42.1% of the students thought that the comments were appropriate and targeted. Another 31.6% of the students thought that the scoring was objective and accurate. For the biggest deficiency of TS, 39.5% of the students thought that correcting the composition was too time-consuming, which inhibited the teacher's enthusiasm in arranging the composition. Another 39.5% of the students thought that the feedback of error information was too slow, which would affect the students' enthusiasm for revision. The vast majority of students (84.2%) think it is possible to combine TS with OAS.

After completing the questionnaire, the research team randomly selected six from the 39 students for interviews in order to better understand the students' feelings on composition correction. The following are the interview questions and the students' feedback.

Interview questions.

1. From your personal experience, talk about your views on OAS.

2. After one semester of OAS and one semester of TS, how do you feel about the two composition feedback methods as a whole?

3. What is your attitude towards the effective integration of the two composition scoring methods? Do you think it is feasible?

Student interview. In order to obtain relatively objective interview results, the research team arranged some students to interview the participants, and then sort the results into written materials based on the recording.

Student A: I think the online automatic scoring speed is very fast, and we can revise it repeatedly according to the error prompt, which can also improve our scores, although there is not much room for improvement. The deficiency is that most of the comments are basically the same, too broad and not very targeted.

For online scoring and teacher scoring, I think they have their own characteristics. Sometimes the teacher gives a score without any written comments. Sometimes there are brief error prompts, and sometimes there is an overall evaluation, but the process is relatively long. It often takes two or three weeks to get the teacher's evaluation results. For the online scoring, we can see the results immediately after we submit it. We will revise it several times to improve our score.

For the integration of the two methods, I haven't considered about it. It should be OK.

Student B: I think the online composition scoring system is very good. It has fast feedback and can be revised repeatedly to improve the score. The error prompt given is also specific and clear. It's just that the comments are a little mechanical.

For the two kinds of scores, I think I prefer online scoring. I can quickly see the evaluation, revise and submit it, and improve my score. Generally, I don't pay much attention to the teacher's feedback. I mainly focus on the score.

For the integration of the two methods, I think it is workable, as it may bring a more objective result.

Student C: My overall feeling is that this online scoring system is still very easy to use, with high efficiency and fast feedback speed. In particular, the students can correct each other's composition. At the beginning, I dare not find fault for the others' composition. I often have to consult the dictionary for confirmation before I find mistakes from the others' composition. After much exercise, I feel I have gained a lot.

On the whole, both scoring methods are OK. Sometimes the teacher will make some encouraging comments on the composition, but the teacher rarely corrects the mistakes. He just underlines the wrong expressions, and sometimes I can't recognize the mistakes.

I think it's a good idea to combine the two scoring methods, so that we can complement the advantages of the two.

Student D: The overall feeling of the online scoring system is good, but it is a little mechanical. If you know its working principle, you can get high scores through some tricks, such as writing more long sentences and making fewer spelling or grammatical mistakes. Since the system will not take the content into consideration, you don't have to consider that too much.

For the two scoring methods, it's hard to say which is better. Each has its own characteristics. The teacher's feedback will be more humanized, with some sentences to encourage the students, but the process of correcting a composition is a little long, while the online scoring is highly efficient.

It would be better if the two scoring methods could be combined.

Student E: The online scoring system is efficient and fast, but it lacks a little personalization. It feels like a machine. It seems that we are dealing with machines and lack emotional communication.

If the two scoring systems are compared, they have their own characteristics. The online scoring is fast and efficient. The teacher gives us comments on the composition at a slow pace, but it is full of warmth. It can give us some words of encouragement.

I think the combination of the two methods is feasible, and can play the effect of making one plus one bigger than two, with both warmth and efficiency.

Student F: The online scoring is not bad, but the comments given are not targeted. We usually revise the composition according to the error prompt to improve our score. We generally don't think much of the comments.

On the whole, the online scoring system is better than teacher scoring. We can improve our score through several revisions, but there is no such opportunity for teacher scoring, because the teacher normally doesn't review our revised composition.

I think the combination of the two is feasible because each has its own characteristics. Combining the two methods will always be more effective than a single method.

Conclusions and Limitations

Conclusions

After the first stage of OAS experiment, it is found that there is a high correlation between OAS and TS (Pearson coefficient being 0.83), indicating that OAS is reliable. After the second stage of the TS experiment and the analysis of the students' questionnaire, it can be seen that the students have a certain degree of recognition for OAS as well as TS, and think that the OAS efficiency is high and the feedback speed is fast, and TS is highly targeted with encouraging comments. Most students (65.8%) think that the two scoring methods have their own characteristics, and the vast majority of students (84.2%) think that if the two scoring methods can be combined, the effect will be better. According to the analysis of the interview results of six students, the conclusion is also consistent with the questionnaire results.

Limitations

This research selected only one experimental class, 39 people, and only 115 compositions. Therefore, the sample size and the number of compositions are not large enough, which may affect the objectivity of the conclusion. In addition, in the questionnaire and interview, only students participate, no teachers, therefore, the information obtained may not be comprehensive enough, which may also be a deficiency of this study. Therefore, a more comprehensive arrangement will be expected for further research in the future.

References

Chen, X., & Ge, S. (2008). Review of automatic composition scoring. *Journal of PLA University of Foreign Languages*, (5), 78-83.

Gao, J. (2021). A study of the rating quality of an automated essay scoring platform: Pigai. *Journal of Harbin University, 42*(7), 102-105.

Ge, S., & Chen, X. (2007a). Exploration of automatic composition scoring for Chinese EFL learners. *Foreign language World*, (5), 43-50.

Ge, S., & Chen, X. (2007b). Research on automatic composition scoring technology abroad. *Technology Enhanced Foreign Language Education*, (117), 25-29.

Ge, S., & Chen, X. (2009). Problems and countermeasures in the research of automatic scoring of college English composition. *Shandong Foreign Language Teaching Journal*, (3), 21-26.

Landauer, T. K., Laham, D., & Foltz, P. W. (2000). The intelligent essay assessor. In K. Hearst (Ed.), *The debate on automated essay scoring. IEEE intelligent systems & their applications* (pp. 27-31). Retrieved on November 12, 2004 from http://que.info-science.uiowa.edu/~light/research/mypapers/autoGradingIEEE.pdf

Li, J. (2011). A case study of teachers' written feedback and students' response in Chinese students' English writing. *Foreign Language World*, (6), 30-39.

Li, Y., & Ge, S. (2008). A study on the validity of graded vocabulary in automatic scoring of college English composition. *Foreign Languages and Their Teaching*, (10), 48-52.

Liang, M. C. (2005). Construction of automatic scoring model for Chinese students' English composition (Doctoral thesis, Nanjing University, 2005).

Liang, M., & Wen, Q. (2007). Review and enlightenment of foreign automatic scoring systems. *Technology Enhanced Foreign Language Education*, (117), 18-24.

Page, E. B. (2003). Project essay grade: PEG. In M. D. Shermis and J. Burstain (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43-55). Mahwah, NJ: Lawrence Erlbaum Associates.

Wu, D., & Zhang, Q. (2011). A comparative study of intelligent and teacher assessment of college English composition. *China Electric Power Education*, (188), 177-178.

Xie, C. (2010). Automatic scoring of English composition and its validity, reliability and operability. *Journal of Jiangxi Normal University (Social Sciences), 43*(2), 136-140.

Zhou, X., Fan, X., Ren, G., & Yang, Y. (2021). Automated English essay scoring method based on multi-level semantic features. *Journal of Computer Applications, 41*(8), 2205-2211.