ISSN 1798-4769
Journal of Language Teaching and Research, Vol. 13, No. 4, pp. 763-771, July 2022
DOI: https://doi.org/10.17507/jltr.1304.09

A Research on the Blended Evaluation Mode in College English Writing Course

Suli Liu
Beijing International Studies University, Beijing, 100024, China

***Abstract*—Evaluation plays a significant role in English writing course, which is an effective way to assess and motivate the learners. Traditional instructor's scoring and feedback is considered helpful, yet straining and not efficient enough. Automated writing evaluation (AWE) is the use of specialized computer programs to grade and evaluate writings in educational settings. The idea of integrating machine feedback with human evaluation is supposed to be comprehensive and efficient. This paper is to investigate the reliability and feasibility of the integration of AWE and human evaluation in college English writing course. An empirical study was conducted in a Beijing foreign languages university, where a blended evaluation mode was applied and the feedback from the students verified that through proper design, the integration of AWE and human scoring was fairly feasible and efficient. The findings of this paper may provide some reference and enlightenment for the evaluation mode in college English writing course.**

***Index Terms*—blended Evaluation, AWE, reliability, English writing**

I. Introduction

It is well acknowledged that evaluation is an indispensable part in English writing course, and is by no means to be underestimated to improve the learners' English writing proficiency. Pedagogically speaking, writing evaluation has always been considered inefficient and time consuming. However, with technological innovation, machine has been introduced into classroom. It has provided more options for assessing and evaluating writing. Thanks to the development of assessing theory and automated writing evaluation, the blended evaluation model of English writing is considered to be a valid approach for writing assessment. In order to upgrade the efficiency of college English writing course, the author carried out a reform on the writing evaluation mode in a Beijing foreign languages university. The tenet of this reform is to combine automated scoring and instructor scoring as a blended evaluation mode to give the learners an efficient and dynamic writing evaluation.

The basic contents of the blended evaluation mode are as follows: iWrite English Writing Assessment and Evaluation System 3.0 (also called iWrite3.0 in the following expressions) was introduced in the formative evaluation in two semesters' English writing course in 2021. The students' writing was evaluated by the integration of iWrite3.0 and the instructor. Students' writing exercises were assigned by the instructor on iWrite3.0 platform, and the students' writing performances (scores, completion time, revising times, habitual mistakes, etc.) were recorded on the platform. Except for grading the students' typical writings, the instructor made a process diagnosis of the students' writing performance. Then the instructor guided the students to carry out targeted exercises, and encouraged the students to complete individual adaptive exercises assigned by the instructor. The course evaluation consisted of the integrated score, in-class performance and individual writing improvement. The purpose of this reform was to improve the efficiency of English writing evaluation and reinforce the objectivity and scientific nature of writing evaluation in the course. This paper is to use qualitative and quantitative research to analyze the effectiveness and feasibility of the above blended evaluation mode in college English writing course.

II. Literature Review

It is widely acknowledged that practice plays a key role in the students' progress in English writing. Research on writing teaching and evaluation shows that not only adequate writing practice, but timely assessment and feedback are critical to improving students' writing ability (Yang& Carless, 2013). In this part, the author is to trace back and analyze the development of writing assessment, AWE, and some empirical studies on the integration of machine and human scoring.

A. Writing Assessment and AWE Abroad

The first attempt on assessing writing dates back to the 1960s, characterized by the publication of *Factors in Judgments of Writing Ability* by Diederich in 1961. Methods used in assessing writing have gone through three major changes. From 1950 to 1970, writing assessment focused on objectivity. The design of structured and limited options was the attempt to minimize possible bias. From 1970 to 1986, holistic and analytic grading became popular. Holistically graded tests focus on the overall quality of the paper while analytically graded ones pay attention to detailed feedbacks. Since 1986, portfolios have gradually replaced timed essays. Scholars believed that portfolio

assessment was more valid because it focused on process rather than product only. In 1994, Aljaafreh & Lantolf found that instructor feedback had positive impact on writing. This finding introduced dynamic assessment into teaching writing.

With technical innovation, machine came to be part of assessing. Automated writing evaluation (AWE), also called Automated essay scoring (AES), is the use of specialized computer programs to grade and evaluate writings in educational settings (Stevenson, 2016). Beginning from 1966, there has developed a couple of well-known AWE programs, namely, Project Essay Grade (PEG), Intelligent Essay Assessor (IEA), "e-rater", Criterion, Writing roadmap.

The appearance of AWE systems has solved some problems in traditional evaluation settings, where human beings, instructor and peers included, were the main raters involved. Traditionally, evaluating took a lot of time and most often, it was not very in time. As people might wonder if AWE would work efficiently, a couple of researches have already proved the effectiveness of AWE on writing development (Li & Zhong, 2017; Wang, 2019; Lee, 2020; Wilson & Roscoe, 2020). Gong et al. (2019) found that with the assistance of human feedback, AWE systems could even work better to promote writing syntactic complexity.

There were also doubts on the practical use of AWE systems. Wang et al. (2015) did not think highly of AWE. They argued that writing was not an isolated tool; after losing its social function, it has broken the connection between writers and readers, resulting in no humanity. Qian et al. (2020) criticized AWE's validity to predict human scores. In face with these doubts, supporters of AWE system made efforts in finding extra approaches to complement the flaws of automated writing assessment. Song (2019) proposed that EFL instructor could work as assistance to automated assessing considering that machine feedback might miss some errors. Li (2019) advocated for multiple evaluations to complement the flaws of computerized feedback. The idea of integrating machine feedback with human evaluation was proved to be more effective and efficient.

B. Domestic AWE Systems

Inspired by the automatic writing evaluation of foreign countries, and combined the practical needs of English writing teaching in China, some domestic experts developed automatic assessment tools as well. Pigaiwang and iWrite are the two leading automated evaluation systems in China. Pigai System is a web-based AWE system, which employed cloud computing and corpus technology. By comparing target compositions with its corpus essays, it analyzes the differences between these two and then gives scores and comments based on the sub-dimensions set beforehand. In 2015, iWrite system, a machine intelligent system for teaching and evaluating English writing, was jointly developed by Foreign Language Teaching and Research Press and National Research Center of Foreign Language Education. This system was designed on the basis of in-depth pedagogical practices. Therefore, it is able to provide feedbacks from four dimensions: language, content, text structure and technical specifications. A big advantage of iWrite system is that it provides coordinated writing teaching function, for instance, the automatic generation of lecture notes or model essays. What's more, it employs joint assessment model, taking account in both machine evaluation and instructor evaluation. This has made it possible for instructor and students to interact with each other.

iWrite system is comparatively new, thus generating only a few studies. The earliest batch of researches on iWrite emerged in 2007. He and Gong (2017) conducted a case study on using iWrite system and summarized the pros and cons of this program, suggesting the combination of intelligent and artificial evaluation.

Liu (2018) trialed on integrating iWrite2.0 system with writing teaching and invested positive results. Later, scholars further pointed out the necessity of integrating automated feedback with other forms of feedback so as to promote writing assessment (Liu & Liu, 2018; Zhou, 2019; Wan, 2020).

C. Empirical Studies on the Integration of Machine and Human Evolution

Although many researchers have acknowledged the effectiveness of integrating human and machine feedback, there are just a few studies made on it. Wu and Zhang (2016) combined instructor feedback with computerized feedback in writing teaching and found that students tended to take more serious instructor' feedbacks, whereas AWE was more like a tool for autonomous learning. Huang & He (2018) suggested that blended feedback was necessary because it could encourage writing revising and promote spontaneous learning. They didn't test the validity of the blended method. Bai and Wang (2019) reviewed the empirical studies from 2000 to 2019, and summarized that AWE system could not replace human feedback and could only be used as a supplementary to instructor evaluation. Chen & Guo (2019) conducted a quantitative study and found that comparing with teach evaluation only, the combination of machine assessment and teach feedback was more effective in writing development. Other scholars have further proved blended evaluation model to be effective in improving students' writing (Gong et al, 2019; Wan, 2020). However, it was found that most of the empirical studies tested only the validity of machine feedback and used only quantitative method. To ensure a scientific and systematic research, the reliability between automated scoring and human scoring should also be paid attention to. Besides, a comprehensive survey on participants is also necessary so as to understand the psychological aspect of artificial and intelligent assessment.

III. Research Design

The current study first tested the reliability between iWrite3.0 scoring and instructor scoring. It then applied

iWrite3.0 system to writing evaluation in an empirical study, combined with instructor assessment and adaptive instructions. At last, it surveyed the participants by questionnaires and interviews. With this, this study gave a comprehensive and scientific view on the reliability and feasibility of the integration of AWE and human evaluation. The participants, research methods, data collection and analysis are introduced specifically in the following part.

A. Participants

The participants are 90 college students majoring in Tourism Management, International Trade, and Finance Management respectively at a Beijing foreign languages university. They are sophomores and have learned English for about 10 years. They are able to understand moderately difficult articles and materials published in the English-speaking countries and express their ideas in general written English. In other words, they are intermediate English learners without specialized writing training. To investigate the feedback, the author also chooses 4 focal students to be interviewed. Detailed data are presented in the research procedure part.

The participants learned college English writing course for 2 semesters in three classes in which a blended evaluation mode was adopted. The objective of this course was to cultivate the students' writing proficiency and polish their writing skills. After the two-semester-learning, the students were expected to be able to write compositions of about 200 words according to the given topic, outline or chart, data, etc. Their writing should be relevant, complete, and well-organized. The students met in the classroom once a week (90 minutes) with the instructor. During this period, they were to learn English writing from the initial diction, sentence writing to paragraph writing and the essay writing. In the first semester, the basic rules and strategies of English writing were emphasized. The second semester focused on "promoting writing by reading" to improve the students' language input and output simultaneously.

B. Blended Evaluation Mode

At the beginning of the experiment, a pilot survey was conducted and the reliability of iWrite3.0 was tested. Above all, we examined the reliability of iWrite3.0 with SPSS 23.0: we randomly selected 30 out of 90 students' compositions. The students' compositions were graded by the instructor and iWrite3.0 respectively. The instructor assessed students' compositions from theme, language, cohesion and coherence according to Outline of National College English Test (2016). Meanwhile, these compositions were evaluated by iWrite3.0. Afterwards, through comparing the scores given by the instructor and those by iWrite3.0, the reliability coefficient was tested by Cronbach's Alpha and Pearson correlation coefficient. With a fairly positive high reliability, iWrite3.0 was approved to be applied in grading and recording students' later writing assignments.

The blended evaluation process went on briefly as follows: the instructor organized the students to learn the concrete English writing rules, strategies and skills in class. Then the students' writing assignments were released by the instructor on iWrite3.0. The scoring system would evaluate the students' compositions from language, content, organization and mechanics. Language part focused on fluency, accuracy and complexity of usage; content referred to relevance and coherence of content; organization emphasized paragraph arrangement and discourse marking; and mechanics referred here to spelling and punctuation. It could also make a process diagnosis for students' writing learning. Thus, the students could get timely grading and revising suggestions of their writing from iWrite3.0. Meanwhile, the instructor would grade one third of the students' compositions and wrote a writing diagnosis for the class based on her own grading and that of iWrite3.0. Then in the following week, the instructor guided all the students to make target revisions based on the feedback from the class writing diagnosis and iWrite3.0. Quite often in this stage, individual adaptive writing exercises were assigned to the students. For instance, exercises to revise run-on sentences or misplaced modifiers, transitional exercises, cohesive exercises, etc. This modification process was often organized in the form of in-class group discussion and later-on revision, for in this case the students can have a thorough understanding on the gradings and many confusions and puzzlements would be clarified and solved. If invited, the instructor would get involved in the discussion, though it's not often the case. Besides grading, there was one more important role iWrite3.0 could play. The students' general performance after class (score of the composition, completion time, revising times, etc.) would be recorded by it. The comprehensive course evaluation of the students was composed of the instructor's evaluation, iWrite3.0's evaluation, the in-class group discussion performance and the final examination (which was graded by the instructor only) according to appropriate weight: formative evaluation accounting for 60%, and final evaluation accounting for 40%.

Hopefully, with the comprehensive and quick feedback, the blended evaluation mode is conducive to boosting the students' writing motivation and helping students timely correct the mistakes in writing.

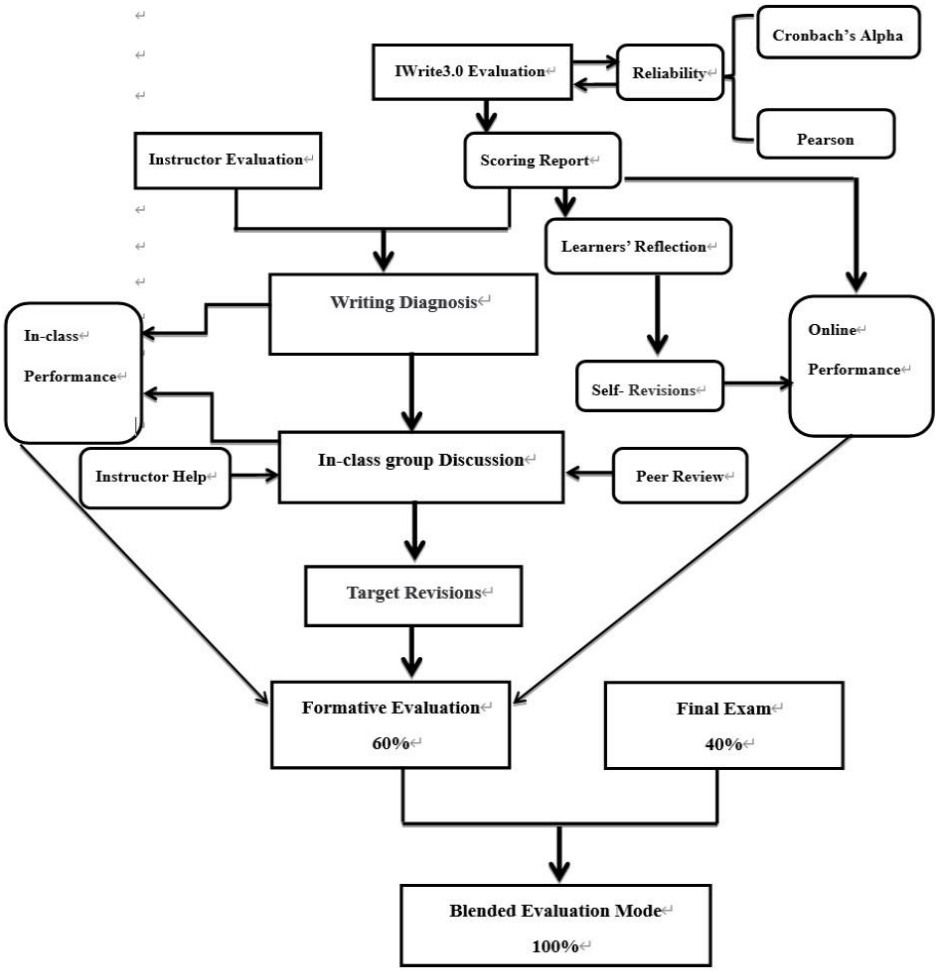

Figure 1 Technical Roadmap of the Blended Evaluation Mode

C. Research Methods

In order to find out the effectiveness and feasibility of the blended evaluation mode in college English writing course, two research methods were adopted in this research. Questionnaire, the instrument of quantitative study, served as one of the main research methods in this study. Besides, this research adopted interviews of the participants as the supplementary research method.

1. Questionnaire

The questionnaire utilized in this research is Satisfaction towards the Blended Evaluation Mode (see Appendix), which is used after this experiment. This satisfaction questionnaire designed by the author consists of four dimensions, which are general attitudes towards the blended evaluation mode, attitudes towards the content of the blended evaluation mode, attitudes towards the instructor and attitudes towards achievement of college English writing course. Each dimension has 5 items. There are 20 items in total. The questionnaire asks the participants to rate on a 5-point scale (1= Very Unsatisfactory, 2= Unsatisfactory, 3= Neutral, 4= Satisfactory, 5= Very Satisfactory). This questionnaire is completed by all the subjects involved. 88 valid questionnaires are collected. The results of these two questionnaires are analyzed by SPSS 23.0.

2. Interview

After the analysis of the results of the satisfaction questionnaire, the qualitative data is collected from the complementary interview with four open-ended questions. Questions in the interview can be divided into four categories: comments on the reliability of iWrite 3.0, comments on the fairness of the blended evaluation mode, comments on the efficiency of the blended mode, and comments on self-achievement in English writing. Before each interview, the researcher asks the interviewee's permission for recording. All the four focal students agree to be recorded. These interviews are recorded and transcribed by the author.

IV. Results and Discussion

The experiment of the blended evaluation mode in college English writing course lasted for two semesters. During this experiment, we analyzed the reliability of iWrite 3.0, Satisfaction towards the Blended Evaluation Mode, and interviews with four focal students. Besides, the students took a model test on CET-6 (College English Test- Band 6) as a routine teaching step.

A. Reliability of iWrite3.0

First of all, the paper analyzed the reliability of iWrite 3.0 by Cronbach's Alpha and Pearson Correlation Coefficient (both belong to consistency estimates) in SPSS23.0. The Reliability analysis in this study refers to the inter-rater reliability analysis, that is, iWrite3.0 is regarded as a grader and the result is compared with those of human raters to determine whether the automated writing evaluation is reliable.

As an index of reliability, Cronbach's Alpha is the most commonly used reliability evaluation tool in social science research at present. Cronbach's alpha is a measure of internal consistency. Generally speaking, the higher the coefficient, the higher the reliability of the tested object. A reliability coefficient of .70 or higher is considered "acceptable" in most social science research situations. In this experiment, the reliability coefficient is shown in table 1.

TABLE 1
CRONBACH'S ALPHA BETWEEN INSTRUCTOR SCORING AND IWRITE 3.0 SCORING

Reliability Statistics	
Cronbach's Alpha	N of Items
0.868	2

The alpha coefficient for the two items is 0.868, suggesting that the items have relatively high internal consistency. The instructor scoring and iWrite 3.0 scoring are highly consistent. That is to say, the internal consistency of the blended evaluation mode is verified.

Pearson correlation coefficient is a linear correlation coefficient, used to reflect the degree of the linear correlation of two variables. The correlation coefficient is expressed as r, and the value is between -1 and 1. According to the degree of relationships, correlation can be classified into the following types: High correlation ($|r| \geq 0.70$), Mid correlation ($0.40 \leq |r| \leq 0.70$) and Low correlation ($|r| \leq 0.40$).

TABLE 2
PEARSON CORRELATION COEFFICIENT BETWEEN INSTRUCTOR SCORING AND IWRITE SCORING

		Instructor Scoring	iWrite Scoring
Instructor Scoring	Pearson Correlation	1	0.789**
	Sig. (2- tailed)		0.00
	N	90	90
iWrite Scoring	Pearson Correlation	0.789**	1
	Sig. (2- tailed)	0.00	
	N	90	90

The reliability of the scores was tested by correlation, and when the number of cases of the variables being examined was greater than 30, the measured results worth being generalized. In this study, 90 writing samples were selected from 3 writing assignments, and each sample was graded by iWrite3.0 and by the instructor. According to table 2, the Pearson correlation coefficient of the two has reached 0.789, indicating that there was a high correlation between these two variables.

In a nutshell，the reliability of iWrite 3.0 was positively verified, and that is to say, the scaler used in the blended evaluation mode is reliable.

B. Students' Feedback on the Blended Evaluation Mode

To learn the students' feedback on the blended evaluation mode in college English course, the author designed and conducted the Satisfaction towards the Blended Evaluation Mode questionnaire. Besides, 4 focal students were interviewed on questions about the blended mode from four dimensions.

1. Questionnaire of Satisfaction toward the Blended Evaluation Mode

The author designed the Satisfaction towards the Blended Evaluation Mode questionnaire. At the end of the experiment, all of the 90 students participating in this experiment were involved in this investigation.

First of all, the reliability is examined by SPSS 23.0 after the satisfaction questionnaires are retrieved from the participants. Reliability is examined by using Cronbach's coefficient alpha. As shown in Table 3, all of these items are above 0.70, which indicates that its reliability is above a commonly acceptable level.

TABLE 3
RELIABILITY STATISTICS ON THE SATISFACTION QUESTIONNAIRE

Cronbach's Alpha	N of Items
.918	20

To have a deeper understanding about students' thoughts on and satisfaction with the practice of the blended

evaluation mode, the author collected and analyzed the questionnaires. The responses are shown in Table 4. All together 88 copies of questionnaires were collected. Here are the results.

TABLE 4
STUDENTS' SATISFACTION TOWARDS THE BLENDED MODE

General Attitudes towards the Mode (n=88)		Attitudes towards the Learning Content of the Mode (n=88)		Attitudes towards the Instructor (n=88)		Attitudes towards the Improvement in English Writing Proficiency (n=88)	
M	SD	M	SD	M	SD	M	SD
4.412	0.307	4.335	0.545	4.489	0.208	4.171	0.882

As it is shown in Table 4, the overall feedback of 88 students is good, for all the means of four dimensions are higher than 4.0, especially the feedbacks to the instructor are pretty positive, which means the practice of the blended evaluation mode is fairly successful. The means of the general attitudes towards the mode and the instructor are higher than 4.4, and the standard deviations of these two dimensions are the lower than those of the other dimensions, which means the students' attitudes are much similar in these two dimensions. However, the participants' feedback on the dimension of improvement in English writing is lower than the other three dimensions, which means the students are not quite sure of their improvements in English writing course. The reasons for this may go into two possibilities: the improvement of English writing proficiency is a progressive, long process, and it's not as apparent as other changes; the students need more comprehensive instructions and practice in English writing. Furthermore, the standard deviation of the improvement dimension is much higher than those of the other three dimensions, which indicates that the students hold quite varied attitudes toward it. The focal students' interviews echo this trend, too. This deviation of the standard deviation of improvement in English writing proficiency shows that improvement is not a uniform activity, and individual attention should be paid on varied subjects. Above all, these findings indicate that the blended evaluation mode in English writing course is feasible and effective.

2. Interview Analysis

To investigate the students' concrete evaluations on the blended evaluation mode in this experiment, the author adopted the semi-structured interview method which mainly focuses on the following dimensions: comments on iWrite3.0, comments on the in-class group work, self-evaluation on English writing proficiency improvement and comments on the blended evaluation mode.

Four open questions are asked in the face-to-face interview by the author:

1. What's your general idea about the introduction of iWrite3.0 in English writing course*?*

2. Do you think the in-class group discussion activities are effective?

3. Do you think this teaching practice helpful to improve your English writing proficiency?

4. What do you think of the blended evaluation mode in English writing course?

Four focal students are interviewed. Here are some of their answers in detail.

The answers of the participants are translated into English and some typical answers are shown as below.

"I think this try of iWrite is necessary and helpful, because I really want a quick and objective response as a contrast of the teacher's evaluation". (S1-Q1)

"I love this system, for it's always there. But to tell you the truth, I prefer the teacher's response". (S4-Q1)

"The in-class group discussion was alright. I could have some help by communicating with my group members and we would have a better understanding of what the system was trying to correct". (S2-Q2)

"I used to prefer study individually, but the group activities gradually aroused my interest by knowing my peer students' brilliant and novel ideas on the topic and I realized content rather than grammar and vocabulary was more valuable in English writing". (S3-Q2)

"I think this teaching practice is very important for me to improve my English writing proficiency. However, there's still a long way to go, for I still have many mistakes in my English writing". (S3-Q3)

"I think my English writing is polished and refined, which may be verified from my grade in the model CET 6, and I'm confident in my coming TOFEL". (S2-Q3)

"I do not regard it necessary to adopt this blended mode all the time, even though it did work in English writing course. I had dreadful experience in my high school oral English AWE evaluation. Only for the proper courses we may have a try". (S4-Q4)

"I love this mode. I think it's a well-balanced one. I want both the teacher's evaluation and the system's, and this mode can give me both". (S1-Q4)

In the English writing course, most of the students' in-class activities took place in group discussion, so it's an objective observation on the students' interactions. To have a clearer picture on the study, the author recorded the four students' performance in group work as well. The students' performances may also verify their ideas in the interview. The students' English level were roughly labeled according to their last semester's final exam and the quiz after this experiment. After the interview, the author rounded up the students' detailed feedback to in-class discussion.

TABLE 5
BASIC INFORMATION OF THE FOUR STUDENTS' INTERVIEWS

	English Level	General Feedback	Performance During the Experiment
Student 1	High	Effective practice	Leads the discussion passively.
Student 2	Medium	Insightful analysis	Always provides opinions.
Student 3	Medium	Novel experience	Gradually participates actively.
Student 4	Low	Quite helpful	Talks much; the most active one.

According to these four students' in-group performance and their comments on the experiment, the author got their feedback to the blended evaluation mode. As for Student 1 with the highest English level among the four students, she's a passive leader in her group for she had answers to many of her peers' puzzlements. However, she gradually realized the valuable thoughts from others and thus the importance of blended evaluation. She suggested that peer evaluation should be involved in this mode as well. Student 2 always provided ideas and explained her understandings of the revisions to her peers, and ways to improve the target writing. Student 3 deemed this blended evaluation mode as a novel experience and was quite satisfied with iWrite 3.0's efficiency. Actually, he volunteered to do more exercises in the scoring system. Student 3 considered this practice an effective way in evoking the students' curiosity: he wanted to know how much progress he's made by writing more and was eager to know the evaluation both from iWrite 3.0 and the instructor. With this inspired writing passion, he got a pretty good mark (13.5 out of 15) in the model CET 6. Student 4 had the lowest English proficiency among these four students. However, he's willingness to communicate inspired the whole group. At the beginning, he's doubtful about iWrite 3.0, not because of suggestions from it, but about the reliability. After three weeks' interactions and the instructor's explanations about the reliability of iWrite 3.0, he started to trust the instant and objective grading. Student 4 deemed the blended evaluation mode a practical and balanced way to be applied in English writing course.

C. Discussion

In this study, the author investigated the design and feasibility of the blended evaluation mode in English writing course. This study mainly focused on the following two main issues: the reliability of iWrite 3.0 and the participants' feedback to this mode. Questionnaires, discussion recordings, weekly writing and interview provided the author with qualitative and quantitative data. The findings are as follows.

Firstly, the design of the blended mode is based on the need analysis of English writing course and the development of automated evaluation systems. In the traditional evaluation mode of English writing, the instructor's feedback period is comparatively long. What's more, instructor's evaluation serving as the only one may inevitably get some subjective factors involved. In the blended evaluation mode, iWrite3.0 or other automated writing evaluation systems can effectively relieve the instructor' workload, saving their limited energy from modifying spelling, punctuation and grammar errors to the guidance of content, the planning and layout of the whole essay, etc. In the research, the blended evaluation mode enabled the participants to get feedback from the objective scoring system anytime and anywhere, in addition to the subjective evaluation of the instructor. Meanwhile, the in-class group discussion could involve their peers' empathy, thus the whole evaluation was instant and comprehensive. As for the duration of the research, many scholars deemed the student's writing proficiency as a dynamic system, which was complex and changed over time. Hou and Chen (2019) argued that the best research time for writing complexity was no less than 1 year. Thus, in this study, the mode was applied in English writing course for 2 semesters in three sophomore classes, and 90 participants were involved to verify the feasibility of this blended evaluation mode.

Secondly, the reliability of iWrite 3.0 was positively verified and the blended evaluation mode obtained supportive feedback from the participants' questionnaires and semi--structured interview. The Cronbach alpha coefficient (0.868) and the Pearson correlation coefficient (0.789) verified the reliability of iWrite 3.0 and the internal consistency of the blended evaluation mode. As for the participants' feedback to this mode, the analyses on the results of the Satisfaction towards the Blended Evaluation Mode questionnaire and the semi-structured interview could fully verify that the participants' feedback to the mode was fairly supportive.

Thirdly, in the research, the author also confirmed some hypotheses about AWE systems. For instance, the author found iWrite3.0 could evaluate language and mechanics accurately, and it quite often paid more attention to some details. In terms of content and organization, the suggestions provided by iWrite3.0 were not enough, and the suggestions were broad and not specific. This finding also verified the necessity of the blended evaluation mode, for AWE system itself was far from enough to comprehensively evaluation the writing.

V. CONCLUSION

The blended evaluation mode of English writing refers to the comprehensive evaluation mode combining various factors, including the instructor's evaluation (in-term and final), the AWE's evaluation and the students' performances in group discussion and their target revisions. It is the specific implementation and practice of the flipped classroom mode. Its purpose is to integrate the advantages of students' collective learning, classroom teaching and network scoring to improve teaching effectiveness and to achieve the "optimal" learning effect.

The average in-term evaluation span from writing assignment to the target revision in this study was one and a half weeks which shortened the previous one by half. The participants could get efficient evaluations from the instructor (even though one third copies each time) and iWrite 3.0 respectively. Moreover, the writing diagnosis (for everyone) was much helpful to many participants, for they deemed the evaluation of iWrite 3.0 was "stamped" and analyzed by the instructor and thus trustworthy. The in-class group discussion served as peer review and could inspire new thoughts and ways of revision. The assessment of the term was composed of the formative evaluation and the achievement test in the final. That is to say, the participants could get overall feedback and comprehensive assessment from the blended evaluation mode.

Actually, the above blended evaluation mode is open to reform. For instance, during the in-class group discussion part, the participants made quite a lot constructive and insightful comments and suggestions to their peers' revisions. In several occasions, the participants also voluntarily voted the best copy of writing in the group and were eager to show it in class. All these indicate that peer evaluation should possibly get involved and account for appropriate weight in future writing evaluation mode.

APPENDIX. SATISFACTION TOWARDS THE BLENDED EVALUATION MODE

Name: Gender: Age:

VU = very unsatisfactory; U = unsatisfactory; N = neutral; S = satisfactory; VS = very satisfactory

		How satisfactory do you find:	VU	U	N	S	VS
General attitudes towards the blended evaluation mode	1	The idea of blended evaluation in college English writing course					
	2	The reliability of iWrite system grading					
	3	The proportion of iWrite system and teachers' grading.					
	4	The fairness of blended evaluation in college English writing course					
	5	The efficiency of the blended evaluation mode					
Attitudes towards the learning content of the blended evaluation mode	6	The learning materials in the course					
	7	The quantity of the assignments in the course					
	8	The quality of the assignments in the course					
	9	The frequency of grading in the course					
	10	The effectiveness of the self-adaptive writing exercises					
Attitudes towards the instructor	11	Your instructor's enthusiasm in the course					
	12	Your instructor's explanation and clarification of the learning target					
	13	The topics your instructor presents in the course					
	14	The encouragement you get from your instructor					
	15	Your instructor's guidance in the class					
Attitudes towards Improvement in English Writing Proficiency	16	Your acquisition of English writing rules					
	17	Your acquisition of English writing strategies					
	18	Your initiative on English writing after the course					
	19	Your confidence in English writing after the course					
	20	The improvement in your English writing proficiency					

REFERENCES

[1] Attali, Y., & Burstein, J. (2006). Automated Essay Scoring With e-rater® V.2. *The Journal of Technology, Learning and Assessment, 4*(3), 3-29.

[2] Bai, L. & Wang, J. (2019). Review of automatic feedback in English composition in recent 20 years. *Foreign Languages Research (01)*, 65-71 + 88.

[3] Chen, L. & Guo, M. (2019). The validity research of AES-assisted teacher's assessment in the era of big data. *Journal of Huainan Normal College (01)*, 80-84.
[4] Gong, W., Zhou, J. & Hu X. (2019). The effects of Automated Writing Evaluation provided by Pigai website on language complexity of non-English majors. *Foreign Language Learning Theory and Practice (04)*, 45-54.
[5] He, Z. & Gong, Y. (2017). Case Study of Application of iWrite English Writing Evaluation System 2.0. *Journal of Chengdu Aeronautic Polytechnic (03)*, 29-32.
[6] Hou, J., & Chen, Z. (2019). A Longitudinal Study on the Development Process of Chinese Engineering Students' English Writing Capacity. *Foreign Languages in China, 16*(03), 63-72.
[7] Huang, J. & He, H. (2018). The impact of the integrated human and AWE feedback on students' writing performance. *Technology Enhanced Foreign Languages (01)*, 19-24.
[8] Lee, Young-Ju. (2020). The Long-Term Effect of Automated Writing Evaluation Feedback on Writing Development. *English Teaching 75*(1), 67-92.
[9] Li, G. (2019). Study on the influence of the multiple feedback on English composition revision. *Foreign Language Education (04)*, 72-76.
[10] Liu, Y., & Liu, J. (2018). Effects of online automated writing evaluation system on EFLlearners' writing revision—An empirical study based on iWrite. *Foreign Language Education in China (Quarterly), (2),* 67-74.
[11] Li, X. & Zhong, L. (2017). Empirical study on automated essay scoring (AES)in college English writing teaching—Based on the pigai system. *Research in Teaching (01),* 57-61.
[12] Qian, L., Zhao, Y., & Cheng, Y. (2020). Evaluating China's Automated Essay Scoring System iWrite. *Journal of Educational Computing Research, 58*(4), 771-790.
[13] Song, Zongwei. (2019). Investigating Chinese EFL College Students' Writing Through the Web-Automatic Writing Evaluation Program. *English Language and Literature Studies, 9*(3). 20-28.
[14] Stevenson, M. (2016). A critical interpretative synthesis: the integration of automated writing evaluation into classroom writing instruction. *Computers & Composition, 42*, 1-16.
[15] Wang, B., Jing, T. & Zhao W. (2015). Automatic writing evaluation research and practice for fifty years— from single, cooperative to interactive. *Foreign Languages Research, 32* (05), 50-56.
[16] Wang, Yusheng. (2019). A Study of Applying Automated Assessment in Teaching College English Writing Based on Juku Correction Network. *iJET, 14*(11), 19-31.
[17] Wan, J. (2020). The application of iWrite2.0 to the blended evaluation of graduate English writing teaching. *Modern communication (06)*, 29-30.
[18] Wilson, J., & Roscoe, R. D. (2020). Automated Writing Evaluation and Feedback: Multiple Metrics of Efficacy. *Journal of Educational Computing Research, 58*(1), 87–125.
[19] Wu, Y. & Zhang, W. (2016). Impact of automated writing evaluation system and teacher feedback on students' writing revision. *Foreign Language Education in China (01),* 12-19 + 91.
[20] Zhou, F. (2019). Research on College English Writing Teaching Mode under the Automatic Writing Evaluation System. *Think Tank Times (27),* 185-186.

Suli Liu is currently Associate Professor in Beijing International Studies University. She earned a Master's degree in TESOL in Dalian Maritime University. She has authored 3 and co-authored 6 books and textbooks in Language teaching. Also she has published more than 20 research papers on Language teaching and Literature. Her main research interests are applied Linguistics and Literature.