Education, Language and Sociology Research
ISSN 2690-3644 (Print) ISSN 2690-3652 (Online)
Vol. 1, No. 1, 2020
www.scholink.org/ojs/index.php/elsr

Original Paper

A Study on the Effectiveness of Automated Essay Marking in the Context of a Blended Learning Course Design

Wenhua Yu[1] & Trevor Barker[1*]

[1] University of Hertfordshire, Hatfield, UK

[*] Trevor Barker, University of Hertfordshire, Hatfield, AL10 9AB, UK

Received: February 7, 2020 Accepted: March 1, 2020 Online Published: April 24, 2020

doi:10.22158/elsr.v1n1p20 URL: http://dx.doi.org/10.22158/elsr.v1n1p20

Abstract

This paper reports on a study undertaken in a Chinese university in order to investigate the effectiveness of an online automated essay marking system in the context of a Blended Learning course design. Two groups of undergraduate learners studying English were required to write essays as part of their normal course. One group had their essays marked by an online automated essay marking and feedback system, the second, control group were marked by a tutor who provided feedback in the normal way. Their essay scores and attitudes to the essay writing tasks were compared. It was found that learners were not disadvantaged by the automated essay marking system. Their mean performance was better ($p<0.01$) than the tutor marked control for seven of the essays and showed no difference for three essays. In no case did the tutor marked essay group score higher than the automated system. Correlations were performed that indicated that for both groups there was a significant improvement in performance ($p<0.05$) over the duration of the course and that there was a significant relationship between essay scores for the groups ($p<0.01$). An investigation of attitude to the automated system as compared to the tutor marked system was more complex. It was found that there was a significant difference in the attitudes of those classified as low and high performers ($p<0.05$). In the discussion these findings are placed in a Blended Learning context.

Keywords

automated marking systems, blended learning, empirical study

1. Introduction

With the continuous development of technology and its use in education, combined with the prevalence of computers and smart devices, Blended Learning (BL) has been integrated into every corner of Higher Education (HE). Learning has radically switched from the traditional mode which was largely reliant on face to face teaching, lectures and textbooks, to multimodal, flexible learning and teaching. BL is a global phenomenon according to Preston et al. (2010) involving a greatly diversified student body. In Chinese universities, it is a requirement that faculties adopt and implement Blended Teaching (BT) in order to meet the diversified needs of the students with the goal of improving the quality of delivery for example as stated by Shanghai Jian Qiao University (2019). There is evidence in the literature that, compared with simply face-to-face and fully online education, a BL approach is beneficial in terms of satisfaction and learning outcomes (Wang et al., 2020; Rupp et al., 2019; Lim & Morris, 2009; Owston, York, & Murtha, 2013). By taking BL into practice, we should make some changes in English education.

1.1 Background to the Study

College English is a compulsory discipline for all majors in colleges and universities in China. College English teaching involves language knowledge, its application, skills, learning strategies and intercultural communication as the main content. It adopts foreign language teaching theories as the guidance, and uses a variety of teaching models and methods. In other words it is a complex teaching and learning system involving the integration of many factors and teaching content, theories, models and methods. One main purpose of the College English course at the university involved in our study is to train students to have strong reading ability, effective listening, speaking, writing and translating abilities in practice. Table 1 depicts the details of abilities to be cultivated by the College English course at Shanghai Jian Qiao University (College English Outline, 2019). The abilities are made up of the following (see Table 1).

Table 1. Abilities Aimed and Cultivated by College English Course

		English abilities					
Autonomous learning	Expression and communication	listening speaking writing reading translation	Responsible and compressive	Collaborative innovation	Service and care	Information application	Global horizon

The course is mainly followed by freshman and sophomore in HE. Another aim of this course is to pursue the socially recognized English certificates like College English Test 4 and 6 (CET 4) (CET 6) which are regarded as essential qualifications in the Chinese job market and are useful as in addition they cover man socio-cultural aspects of English. English test marks are important for learners in the Chinese job market.

1.2 Details of Teaching English Writing Practice

Several changes in teaching were made in the Jian Qiao University in order to adapt to the perceived need of teaching English writing. Among the four basic English skills of listening, speaking, reading and writing, improving English writing ability has always been a difficult task confronted by Chinese teachers and students (Sun, 2014). In College English teaching at Jian Qiao University, the number of students in each class is large. This puts forward higher requirements. Teachers are required to spend inordinate time and energy marking students' compositions. In a single semester, teachers were only able to assign one or two writing task for that reasons. Consequently, there were very few opportunities for the students to practice their writing and consequently for teachers to review progress and provide feedback.

In the experience of one of the authors of this paper, teaching English writing raises many problems related to the process of grading students' English composition. Firstly, it requires a considerable amount of time to grade essays and provide useful, timely and relevant feedback and evaluation to individual students. Secondly, grading can often be subjective when scoring students' writing. There is a possibility that students may be stereotyped according to scores obtained rather than their individual strengths and weaknesses. It is possible that demographic factors such as gender, age, ethnicity. prior performance on tests and other courses and socio-cultural factors may conceivably influence feedback and in extreme cases, the grade obtained by learners. It is often difficult for a teacher to be entirely neutral in their approach to marking essays. For these reasons scoring essays becomes an enormously complex cognitive task that involves a multitude of inferences, choices, and preferences on the part of the grader. The exact features are attended to in an essay, the characteristics and sections that are weighted most highly, and the standards adhered to are all factors that may vary widely across human graders. Indeed, it has been observed that teachers' ratings of essays can be highly variably and often not objective (Huot, 1990; Huot, 1996; Meadows & Billington, 2005).

Additionally, the class size in Chinese universities is often very large. A teacher may often teach a class with more than 50 students. If he or she teaches several classes in parallel in one semester, then he or she is required to grade several hundred essays. Consequently, essay rating becomes an arduous task for teachers. Teachers often devote a great deal of effort, many students appear only to be concerned with the final score and less so with the feedback and feed-forward provided by the teacher. Students may be unwilling to review and reflect the feedback or evaluation from the teacher. This factor makes it difficult to help a student to improve their writing prior to the next task. A possible reason for this may be the timeliness of the feedback. Fast, efficient feedback is likely to ensure that help is provided

in good time to assist in the next task. If feedback is too slow, then student are likely to pay less attention to it we argue.

Plagiarism is a growing concern in universities across the globe. The prevalence of electronic resources, copy and paste and file sharing has made it easy for some students to cheat. Manual marking of essays is slow and complex as described above. It is therefore difficult to detect plagiarism on students' paper-based writing. The grading of English writing effectively and to provide useful, timely and effective feedback in a timely manner becomes an important task.

Against this background, in the context of a BL design, Automated Essay Scoring (AES) online has been adopted at the Shanghai Jian Qiao university. AES is defined as a computer technology that is able to evaluate and grade written works (Shermis & Burstein, 2003). At the Shanghai Jian Qiao university, the English writing course is delivered by face to face lectures and tutorials in classrooms. And an online system of AES is has been implemented to supplement the traditional classroom teaching. Using technology to supplement the real classroom teaching is a fundamental objective of China's foreign language teaching as explicitly stated in the National Curriculum of College English Course (2017). According to Kaleta et al. (2007), teachers who design BL courses often place additional online elements within a traditional course framework without removing current activities. This phenomenon is also referred to as "the course-and-a-half syndrome" (ibid., p. 127).

Figure 1 below summarizes the type of BL design employed within this study. Instruction is delivered in the classroom while all the necessary exercise and practice are completed online after class is over. This may be considered as a basic way of combining traditional classroom teaching with supplemented web-based activities. Many instructors design BL courses in this way according to several researchers, for example (Brunner, 2006; Kaleta et al., 2007). The addition of extra activities to an existing, traditional course as employed in this study may be referred to as a basic-level blend.

Figure 1 illustrates applying the basic-level blend approach to English writing course design. Then this leads to the research objectives of this study.

- How to test the effectiveness of this basic-level blend?
- What are the advantages and disadvantages of this basic-level blend?

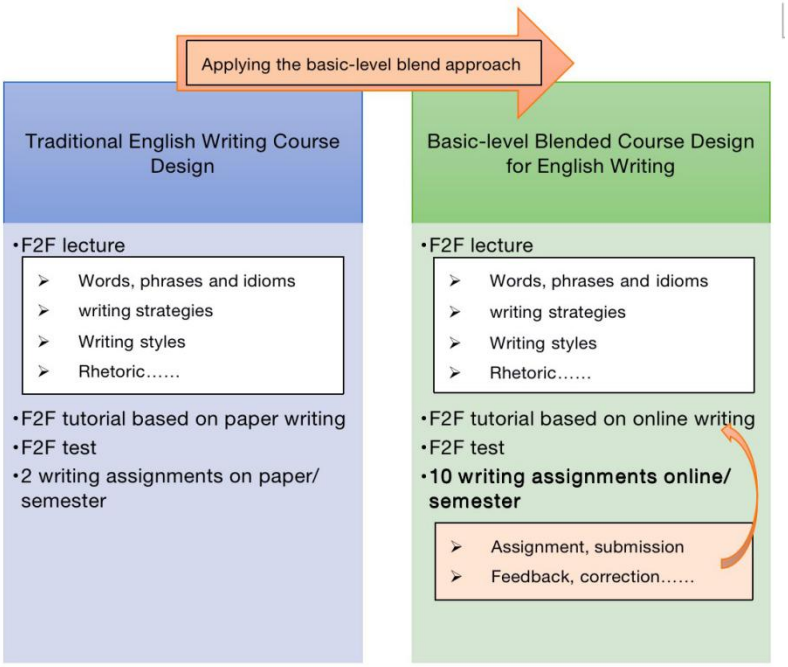

Figure 1. Basic-Level Blended English Writing Course

2. Literature Review

In this brief literature review, four main areas relate to the context of our research and are covered as follows:

① How is AES developed?

② What are the claimed benefits and claimed limitations of AES?

③ How do teachers perceive BL and how does the perception impact the course design?

④ What is the attitude of teachers to basic-level blended course design for English writing ?

2.1 A Brief Review of Studies on AES

More than 50 years ago, Ellis page (1966) predicted the arrival of the so called “teacher’s helper”, that would grade papers by computer (Shermis, 2014). Just seven years later, Page and his colleagues at the University of Connecticut developed the first automatic essay grading engine, which was called Project Essay Grade (PEG) (Ajay, Tillett, & Page, 1973; Shermis, 2014). For reasons related to the difficulty of entering text within this technology the system did not gain immediate popularity until the early 1990s. From then on, some commercial and also several non-profit organizations took up exploring different types of essay scoring systems for English language. AES systems at that time were adopted by testing companies, universities, and public schools (Toranj & Ansari, 2012). The most widely known AES systems include Project Essay Grader (Page, 1966, 1968, 2003), the Intelligent Essay Assessor (IEA; Landauer, Laham, & Foltz, 2003), CriterionSM, e-rater (Attali & Burstein, 2006; Burstein, 2003), and IntelliMetric (Rudner, Garcia, & Welch, 2006), MY Access® and BETSY (Toranj & Ansari, 2012). For reasons outlined in the introductory section above AES system development

became commercially competitive around this time as it was able to combine the teaching of English writing and the development of large-scale tests of writing. Two major commercial organizations in the United States of America with significant financial support from the government, promoted AES as an acceptable scoring mechanism. AES was put forward as a viable tool for evaluating students' performances in some important large-scale tests such as GMAT, GRE, and TOEFL (Bay-Borelli et al., 2010).

In the general literature related to AES, the evaluation process for AES covers a number of criteria, including association with human scores, distribution differences, subgroup differences, and association with external variables of interest (Ramineni & Williamson, 2013). Such testing is essential to establish the validity of AES systems. Teachers have to be confident in the reliability and validity of AES systems and also be aware of the limitations of AES. A major issue in the research presented in our paper related to the possibility of improving students' scores. This might be expected if the improved feedback and extra activities were of benefit. The results of other researchers in this area are uncertain, for example (Wilson & Roscoe, 2020).

Studies on AES systems have demonstrated that computers can function as more effective cognitive tools (Joundy et al., 2019; Attali, 2004; Toranj & Ansari, 2012). Researchers have found that the AES system could be useful as it was able to give scores and feedback to students rapidly (Page, 2003). Previous studies have shown that high correlation can be achieved between manual scoring system and AES system (Kukich, 2000; Attali & Burstein, 2006; Ben-Simon & Bennett, 2007; Toranj & Ansari, 2012).

Some scholars have compared AES with human raters. According to Shermis (2014), AES performed well in five of the seven tests and was close to human raters in the other two. Further studies on the validity of AES systems, have suggested that they are able to play a practical role in the assessment of high-risk writing (Shermis, 2014).

Alignment with human scores on essays should not be the only validity criterion according to Wilson and colleagues (2020) Bennett and Bejar (1997) and Bennett and Zhang (2016) Sara Cushing Weigle (2013) notices the significance of systematically articulate the complicated structure of second language writing instruction and evaluation in her book English Language Learners and Automated Scoring of Essays. It was reported that AES is more consistent across multiple assignments in comparison to human raters. However, as stated in her paper, the operational rules of AES are not able to capture the characteristics of non-native writing. Human raters are sensitive to these more specific characteristics when marking the essays. Her conclusion from her research with English learners studying a foreign language emphasizes the need to understand the students' diverse needs in the first place, first when system developers are designing AES systems. It is also important for teachers when they are developing courses that include additional activities from AES. The more they know about the students' needs, the greater the possibility of satisfying the diverse needs of an increasingly larger population (Weigle, 2013; Elliot & Williamson, 2013).

However, because writing is an activity that is so deeply human, its association with formulation is double edged (Elliot & Williamson, 2013). Because students are encouraged to write fluently or to achieve competency in their knowledge of conventions, a certain degree of formulation is necessary (Elliot & Williamson, 2013). But when these formulations are used by machines as the basis for assessing writing beyond fluency or knowledge of grammar (Attali & Powers, 2008) there is an inherent suspicion that technology can corrupt the essence of a fundamentally human activity (Ericsson & Haswell, 2006; Herrington & Moran, 2012).

2.2 Self-Efficacy

Gairs (2007) showed that some students had a higher satisfaction rating with online learning systems though they did not necessarily have their performance enhanced or behavior changed by the use of AES systems. This was attributed not to the use of AES system per se, but to their willingness to an inherent engagement with such systems. Motivational processes such as reflection and self-efficacy were likely to be responsible to the high attitude scores it was postulated. Researchers have argued that it was necessary for learners to take part in the reflective activities if it were to result in a significant improvement in self-efficacy and task value in online activities (Qian et al., 2019). Self-reflection may be improved by a constructive BL approach in which the students assess their own work based on feedback and a knowledge of assessment criteria in relation to their individual performances and goals. Learners may then have affective cognitive reactions guided by their self-judgments and might be able to make decisions based on previous learning and hopefully relate this to future tasks and goals. It is hoped that this hypothesized effect may be measured by an increase in self-efficacy at the end of our study.

Efficacy emphasizes the ability and confidence to achieve a goal satisfactorily. It relates to one's belief in a capability to perform a specific task. It determines how people feel, think, motivate themselves, and it also refers to their confidence to achieve the desired outcome (Bandura, 1986). Individuals' task-specific self-efficacy can be generalized to a wide range of tasks or activities in certain disciplines (Bandura, 1997). Bong (2001) found that students' self-efficacy judgments contain strong subject-specific components. A variety of studies have revealed the role of self-efficacy in a range of disciplines and contexts, from elementary school mathematics (Phan & Walker, 2000), computer-based science learning (Liu, Hsieh, Cho, & Schallet, 2006), and writing (Pajares & Valiante, 1999), indicating that that students' self-efficacy is an important factor in predicting their learning performance or achievement. Self-efficacy it may be argued, mediates people's interpretation of their knowledge, skills, or experiences of prior attainments, and is believed to be an essential factor in positively predicting learning outcomes. According to Bandura, students' learning experiences play an important role in explaining their self-efficacy of learning (1997). In our research the use of AES an a BL context is predicted to increase the self-efficacy of learners.

2.3 Curriculum Added with AES

A model that has empirically been demonstrated to yield substantial gains for students was described in the book "The Framework for Success in Postsecondary Writing" (CWPA, NCTE, NWP, 2011) and also by Graham and Perrin (2007). The general purpose of the study presented here is to explore the advantages and disadvantages of a basic-level blend with AES. It is hoped that this may help teachers to have a deeper conception of BL in a real context and to help students improve their English writing experiences. This will involve the learning of phrases, idioms, writing styles, skills, conventions, strategies, rhetorical knowledge and critical thinking.

2.4 Details of the Online AES Software Used in This Study

This AES system used claims that it is able to provide timely, comprehensive and effective grades and diagnostic feedback to students' writing online. It is claimed that it is able to enable students to understand better their own English composition, to correct mistakes themselves in time in order to improve their English ability. Teachers are also able to assess the overall writing level of students, in order to conduct targeted tutorials for learners, based on their performances. With the help of this system's automatic review, teachers would be able to arrange more pertinent writing assignments easily, thus effectively solving the traditional teaching problem "students are unwilling to write, while teachers are unwilling to mark" (AES online, 2019).

It is also claimed by developers of the system that the system can analyze a composition from the aspects of spelling, content, organization, word choice and grammar, providing multidimensional personalized feedback information, which can be used for formative and terminal evaluation of the students. It can play an extremely important role in improving students' language ability (AES online, 2019). To sum up, this AES System is claimed to function in support of the following traits:

- High credibility of the score
- Strong ability of diagnosis and error correction
- Featured detection function (AES online, 2019)

This study intends to test the effectiveness of the basic-level blend in the course design of English writing by adding extra activities online without eliminating any traditional on-class activities. Then the advantages and disadvantages of this approach can be analyzed and identified. Here followed research questions driven by the research objectives.

- How to test the effectiveness of this basic-level blend?

  ① Can we observe any significant differences in performance between students using basic level blend approach adding system and students using traditional method only with paper-based practice?

  ② What is the relationship between learning outcomes and learners' satisfaction with the experience from this basic-blend?

- What are the advantages and disadvantages of this basic-level blend?

① What factors should be considered by teachers in HE when they choose this basic-level blended course design?

② What can be improved in this basic-level blend?

3. Method

3.1 Participants

The experiment involved two groups of learners who were required to produce eleven essays. One group assessed and given feedback by tutors (the control group) and the second group using the online AES system. Participants were 2 groups of undergraduates from non-English majors in a Chinese university. Groups were balanced as far as possible in the context of an ex-post facto study. The demographic variables are shown in Table 2 below. Both groups consisted of similar number of male and female undergraduates aging from 18-19. Both groups classified as having achieved intermediate level according to their English proficiency on entry to the university. Groups were selected using a quasi-random sampling strategy.

Table 2. Details of Participants in the Study

	Tutor marked participants	AES marked participants
N	36	35
Mean age (years)	18,8	18,3
Gender (F/M)	16/20	16/19
Academic English level	CET4	CET4

According to proposed by Zimmerman (2002) there are three stages of self-regulated learning strategy. These include forethought, performance, and reflection. Learners in this study were required to complete these three stages in their course. Students set learning goals prior to starting a task in the forethought stage. Students then engaged in and completed an essay writing task (performance). Feedback provided was intended to allow students to reflect on the learning process. How self-regulated learning strategy was employed in this study is explained below.

Forethought: The students in both groups were given an orientation about the course by the tutor, including the conception of feedback, evaluation, goal setting, writing instructions and reflection. For the experimental group, the teacher also demonstrated how to use AES system. The students acclimatized themselves to the feedback and evaluation mechanisms in the AES system. For the control group, the tutor demonstrated simple administrative procedures such as submitting work, how to make corrections according to the feedback and evaluation from the teacher and how to store their work. These were functions achieved fairly simply in the online system.

Performance: The duration of the experiment was approximately 17 weeks, and Table 3 presents the essay topics that were assigned to both groups.

Reflection: After completion, students reflected on their learning processes either through the writing feedback and evaluation mechanism provided by AES or from the teacher's paper-based comments. Reflection then related to the amount and quality of feedback given to participants by the tutor and online system. Although this was not directly assessed in this research, learner attitude to the process was measured which was assumed to relate to learners' reflections of the experience.

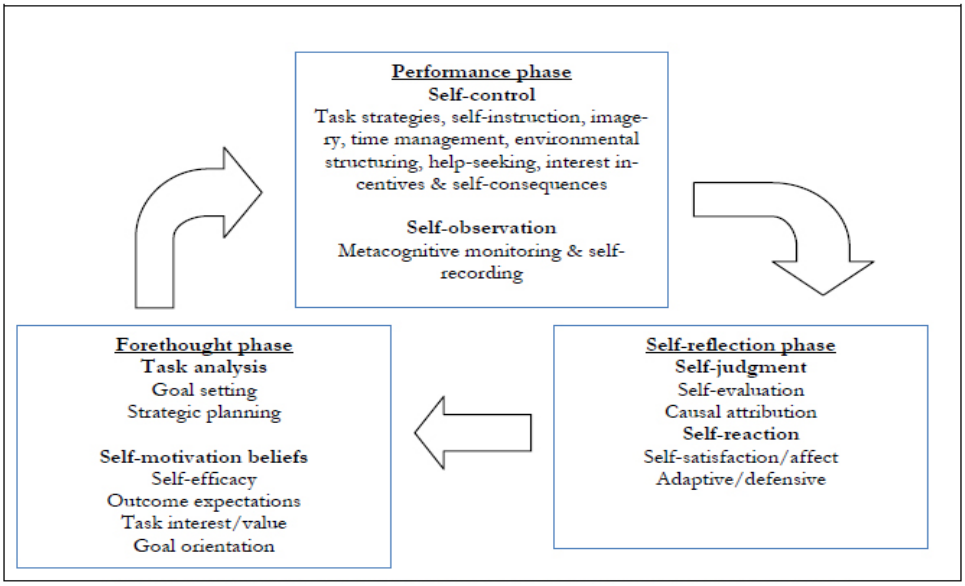

Figure 2. Phases and Processes of Self-Regulation according to Zimmerman and Moylan (2009)

At the end of the study, students were asked to rate their perceived difficulty of each of the essay topics on a 10-point Likert scale, and also to complete a short questionnaire on their experience of and attitude to English language essay writing.

The students were assigned writing tasks respectively online and on paper every 10 days throughout the duration of the study. The topics (as shown in Table 3) were selected from the **CET 4** category from the AES system under investigation. Each essay set clear requirements on the length and structure for both groups of participants.

Table 3. Topics Assigned

number	Topics
premeasure	Write a letter apologizing for being late for
13455	Why I Chose the Major of …
13457	My Favorite City in China
13868	True Friendship among Roommates
13945	Lucky Money
14222	The Advantages of Getting a Good Education
14359	Should we go after fame and fortune?
14445	Part-time Job in This Summer Vacation
14446	An Unforgettable Party
14449	On College English Teaching
14594	Textbook Knowledge or Social Skills

In order to avoid the Hawthorne effect (Levitt & List, 2011), students in the experimental group were not informed of the experiment, and the experiment was naturally integrated into this basic-level blended course. In order to avoid the John Henry effect (Saretsky, 1972), students in the control group were not informed of the experiment, either. Both groups were taught by the same teacher, and received the same curricular content, teaching schedule, requirements, and goal setting.

Table 4. Activities Undertaken by the Experimental and Control Group

Activity	Experimental	Control Group
self-regulated composition	Online	Paper
Feedback and evaluation	Several times	Once on Paper
Reflect and review	Several times	Once on Paper
Correction, editing and	Several times	Once on Paper
Archive of material	Online	Paper
Repeating the previous	Yes	No
Estimation of difficulty	Yes	Yes
Questionnaire on efficacy	Yes	Yes

4. Results

A comparison between paper essay and the Panorama online system was undertaken as previously described. A pre-test was completed by both groups to test if there were differences in the samples. The results of this are shown in Table 5 below.

Table 5. The Results of Pre-Test between Participants in Online and Paper-Based Essay Marking System

Group	N	Mean	SD
Online	35	71.1	9.2
Paper	36	68.2	7.1

In order to test the significance oif any difference in the means of the two groups, an independent samples t test was performed. The results of this test confirmed that there was no significant difference between the mean performances of the groups (t=1.45, df=69, p=0.15). It was noted that although there was no significant difference in the means, the online students exhibited a slightly higher mean score than the paper based students.

A comparison was made between the performance of the students as they undertook 10 essay assignments. The mean results of the essays and their topics are presented in Table 6 below.

Table 6. A Comparison between Online and Paper-Based Mean Essay Scores

Essay	Topics	Mean Score (Paper-Based)	Mean Score (Online)
Pre-Test (E1)	Write a letter apologizing for being late for class	68.24	71.08
Essay 2 (E2)	Why I Chose the Major of …	64.36	70.71
Essay 3 E3)	My Favorite City in China	63.39	68.86
Essay 4 (E4)	True Friendship among Roommates	68.36	74.86
Essay 5 (E5)	Lucky Money	63.86	69.06
Essay 6 (E6)	The Advantages of Getting a Good Education	70.49	73.04
Essay 7 (E7)	Should we go after fame and fortune	67.67	72.88
Essay 8 (E8)	Part-time Job in This Summer Vacation	70.26	78.13
Essay 9 (E9)	An Unforgettable Party	69.77	73.44

Essay 10 (E10)	On College English Teaching	71.89	75.40
Essay 11 (E11)	Textbook Knowledge or Social Skills	72.59	75.57
Mean Total		68.26	73.19

In order to obtain an informal understanding of the performances of the two groups, a graph was plotted showing how the performance of the two groups varied with time. This is shown in Figure 3 below.

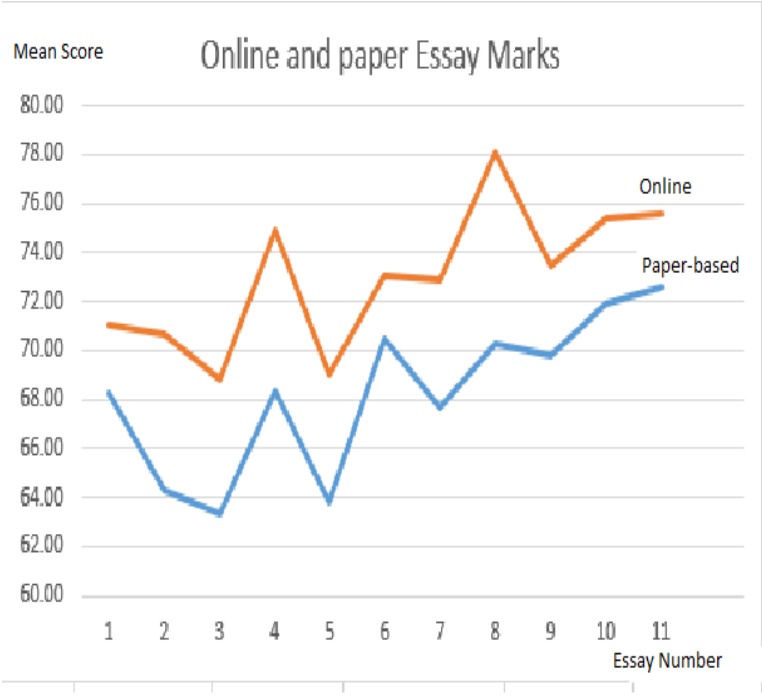

Figure 3. Graph of the Comparison between Online and Paper-Based Mean Test Scores

It is interesting to note that the shape of the curves is similar. In general essay scores in the tutor marked system corresponded to those in the online automated system. To further understand any between the performances of the groups in the essay assignments, a 2x10 mixed ANOVA was performed on the data summarized in Table 6 above. The results of this ANOVA were ($F=9.845$, $df=1$, $p=0.003$). The value of ($p<0.01$) compels us to conclude that there was a significant difference in test scores between the online and paper-based groups. The mean values of the test scores (from Table 1 above) were Tutor marked=68.26; Online=73.19. We are able to conclude on average, the online automated system learners performed better that the control group.

A post hoc analysis was performed on the data summarized in table 6. The results of an independent ANOVA are shown in Table 7 below.

Table 7. The Results of an Independent ANOVA on the Means of the Paper-Based and Online Conditions

Essay	Mean score Tutor based	Mean score Online Automatic	df	F	P (one-tailed)
E2	64.36	70.71	70	6.207	0.08
E3	63.39	68.86	70	4.558	0.02
E4	68.36	74.86	70	10.016	0.001
E5	63.86	69.06	70	5.457	0.01
E6	70.49	73.04	70	1.471	0.12
E7	67.67	72.88	70	7.852	0.004
E8	70.26	78.13	70	13.428	0.000
E9	69.77	73.44	70	2.648	0.050
E10	71.89	75.40	70	1.899	0.09
E11	72.59	75.57	70	1.593	0.10

It is evident that essays E2, E3, E4, E5, E7, E8 and E9 had significant differences in performance between paper-based and online conditions (p one-tailed<0.05). Possible reasons for the lack of a significant difference in essays E6, E10 and E11 (p>0.05) will be discussed later.

The overall shape of the graph presented in Figure 3 is interesting. It suggests that both groups had improvement in their scores over time. This is important as it suggests that the paper-based and online systems were both effective in improving the performance of learners. In order to test this hypothesis, a Pearson's PM correlation was performed on both groups in order to test the significance of any correlation between the test scores and study time.

The output from this correlation is presented below in Table 8.

Table 8. Correlation between Mean Essay Scores for Paper-Based and Online Conditions and Study Time

		Paper based	Online automated	Study time (weeks)
Tutor marked	Pearson Correlation r	-	0.839	0.743
	Sig. (1-tailed)		0.000	0.005
Online automated	Pearson Correlation r	0.839	-	0.680
	Sig. (1-tailed)	0.005		0.011
Study time (weeks)	Pearson Correlation r	0.743	0.680	-
	Sig. (1-tailed)	.0005	.0011	
N		11	11	11

Significant positive correlations were found for both paper-based essays ($r=0.839$, $p<0.001$) and online essays ($r=0.680$, $p=0.011$) with study time. This suggests that there was a significant positive relationship between study time and essay score for paper-based and online essays, showing that scores improved over the duration of the course.

The results suggest that in both cases learners improved in their scores over time and that the performance of learners on the essays were also related. This is an important finding in the context of this research. It is important to show that learners are not disadvantaged by a new intervention. We can conclude that the online system is at least as effective as the traditional paper-based system at supporting learners in their essay writing.

The overall shape of the graph displayed as figure one is also interesting as the shape of both curves is similar, which supports the above finding.

In order to explore more fully the shape of the graph in Figure 1 above, a further investigation was performed. Learners ranked their perceived level of difficulty for each essay on a Likert scale (1 to 10) where 1 easy and 10 is difficult. It would then be possible to investigate any relationship between perceived difficulty level and the scores obtained in the essays. A summary is presented in Table 9 below.

Table 9. Perceived Difficulty Levels for Essay for Tutor Marked and Online Automated Systems with Scores

Essay	Mean perceived difficulty (online)	Mean perceived difficulty (tutor)	Overall Mean	Mean Essay score (online)	Mean Essay score (tutor)	Overall Mean
E1 (pre-test)	4.43	4.29	4.36	71.08	68.24	69.66
E2	3.43	3.57	3.50	70.71	64.36	67.54
E3	2.00	2.07	2.04	68.86	63.39	66.13
E4	6.36	4.57	5.46	74.86	68.36	71.61
E5	4.43	3.07	3.75	69.06	63.86	66.46
E6	6.36	7.00	6.68	73.04	70.49	71.77
E7	6.00	6.79	6.39	72.88	67.67	70.28
E8	7.50	6.57	7.04	78.13	70.26	74.20
E9	5.21	7.07	6.14	73.44	69.77	71.61
E10	6.93	7.29	7.11	75.4	71.89	73.65
E11	8.14	8.49	8.32	75.57	72.59	74.08

It is interesting to note that those essays (E6, E10 and E11) from Table 3 above, where there was no significant difference in performance between the two groups, had relatively high perceived difficulty levels. This factor may account for the lack of a significant difference. The relatively low level of alpha for these exceptions, in the region of (p=0.1) suggests that despite a lack of significance there may still be a slight positive effect.

In order to test any significance in the relationship between difficulty ratings and performance, a Spearman's correlation was performed on the data summarized in Table 5 above. The results of this correlation are presented in Table 10 below.

Table 10. Relationship between Perceived Difficulty Level and Essay Scores for Online Automated Essays and Tutor Marked Essays and Mean Scores

		Online marked rating	Tutor marked rating	Mean rating	Online marked score	Tutor marked score	Mean score
Online marked rating	Coef. (rho)	1.000	0.664	0.891	0.891	0.800	0.870
	Sig (1-tailed)		0.013	0.000	0.000	0.002	0.000
Tutor marked rating	Coef. (rho)	0.664	1.000	0.873	0.782	0.909	0.820
	Sig (1-tailed)	0.013	-	0.000	0.002	0.000	0.001
Mean rating	Coef. (rho)	0.891	0.870	1.000	0.882	0.882	0.920
	Sig (1-tailed)	0.000	0.000	-	0.000	0.000	0.000
Online score	Coef. (rho)	0.891	0.782	0.882	1.000	0.882	0.970
	Sig (1-tailed)	0.000	0.002	0.000	-	0.000	0.000
Tutor marked score	Coef. (rho)	0.800	0.909	0.882	0.882	1.000	0.934
	Sig (1-tailed)	0.002	0.000	0.000	0.000	-	0.000
Mean score	Coef. (rho)	0.870	0.820	0.920	0.970	0.934	1.000
	Sig (1-tailed)	0.000	0.001	0.000	0.000	-.000	-
N		11	11	11	11	11	11

The results of the Spearman's correlation shown in Table 6 above were highly significant at (p one-tailed<0.001) in most cases. This was taken to indicate that the test scores were indeed positively

correlated with perceived difficulty level. Also there was a significant relationship between the perceived difficulty level of online automated marked essays and tutor marked essays (p one-tailed=0.013).

A Mann Whitney U test was performed to test the significance of any difference in the ranking between online automated and tutor marked essays. The results of this analysis showed that there was no significant difference between the perceived difficulty level of the two groups (N=11, U=51.00, p=0.533).

An attitude questionnaire was administered to the participants in order to investigate any relationship between performance and attitude to the essays. The results of the questionnaire are shown in Figure 4 below (based on a Likert scale where 1 represents a negative attitude or opinion and 5 a positive one).

	Means	mean online	mean paper
Q1: GENDER	Q1	1.59	1.5
Q2: My average score for writing assignments this semester.	Q2	3.5	2.86
Q3: I think English writing is a very important part of English learning.	G3	4.41	4.32
Q4: i am confident that my English writing is very good.	Q4	3.18	2.64
Q5: I can write according to requirements of the writing task	Q5	3.91	3.77
Q6: I can spell the words correctly.	Q6	3.64	2.82
Q7: I can use the punctuation correctly	Q7	4.23	3.64
Q8: I can correctly use different properties of words, such as noun, verb, adjective, etc.	Q8	3.91	3.36
Q9: I can write compound and complex sentences with proper punctuation and correct ...	Q9	3.64	3.05
Q10: I can combine several sentences into one paragraph to clearly express a topic.	Q10	3.91	3.41
Q11: I was able to write a well-structured article that expr...	Q11	3.86	3.23
Q12: i can use different writing techniques properly in my writing.	Q12	3.55	3.18
Q13: I feel like making some corrections according to the feedback given to my composition	Q13	3.64	3.68
Q14: I evaluate my writing progress to see if I achieve the goals set by myself	Q14	3.73	3.27
Q15: I often reflect and review my essays stored in the system to	Q15	3.64	3.18
Q16: I am satisfied with my current writing progress.	Q16	3.59	3.18
Q17: I am engaged in my current writing progress	Q17	3.77	3.23
Q18: I have motivation to write	Q18	3.68	3.14
Q19: I enjoyed the whole writing process.	Q19	3.74	3.25

Figure 4. Results of an Attitude Questionnaire for Online Automated and Tutor Marked Groups

A Mann Whitney U test was performed to test any difference between the mean attitude and mean essay score for online automated and tutor-marked essays. The results of this analysis suggested that there was a significant difference between the attitude of learners in the online automated and tutor marked essay groups. (N=19, Mean rank Online=25,39, Tutor marked=13.61; U=68.5, p (one-tailed)=0.001). The mean ranking shows that the learners with online automated essay marking rated higher than those with tutor marked essays.

A correlation was performed to investigate the significance of any relationship between essay score and attitude. Figure 5 below shows mean essay scores and attitude for the two groups of learners.

StuPaper	PaperEssay	PaperQuest	StuOnline	OnlineEssay	OnlineQuest
1721107	63.00	1.70	1722120	75.00	3.80
1721108	68.80	3.50	1722122	68.20	3.50
1721109	70.50	3.30	1722123	69.00	3.60
1721113	49.30	3.30	1722128	73.10	3.80
1721116	76.00	3.30	1722130	69.50	3.30
1721117	67.40	1.70	1722132	78.20	3.30
1721119	59.40	2.90	1722133	68.10	4.10
1721121	72.00	3.20	1722134	62.30	3.30
1721125	66.90	3.90	1722136	68.80	4.00
1721126	69.50	3.70	1722138	81.70	2.80
1721126	63.10	3.20	1722139	78.70	4.00
1721128	71.00	2.80	1722141	73.10	2.80
1721130	74.70	4.00	1722142	72.30	4.70
1721133	80.00	3.70	1722143	80.40	4.70
1721135	67.80	3.40	1722144	86.00	3.80
1722079	76.60	3.90	1722149	66.40	3.60
1722081	72.80	4.00	1722153	76.80	4.10
1722096	68.00	3.20	1722155	80.90	4.10
1722099	78.00	3.10	1722156	87.10	3.40
1722100	72.80	3.50	1722158	81.50	4.90
1722102	73.00	3.60	1723965	62.40	3.10
1723982	65.10	3.50	1723972	71.50	3.70

Figure 5. Mean Essay Scores and Attitude Scores for the Online Automated and Tutor Marked Groups

The results of a Spearman's correlation on the data displayed in Figure 5 are shown in Table 11 below.

Table 11. Results of a Spearman's Correlation between Mean Attitude Scores and Performance for Two Groups of Learners

	Tutor marked Essay Score	Online marked essay score	Tutor marked attitude score	Online marked attitude score
Tutor marked essay score (rho)	1.000	0.023	0.0416	-0.099
p(one tailed)	-	0.460	0.027	0.330
Online marked essay score	0.023	1.000	-0.127	0.231
p(one tailed)	0.460	-	0.287	0.098
Tutor marker attitude score	0.416	-0.127	1.000	0.287
p(one tailed)	0.027	0,287	-	0.098
Online marker attitude score	-0.099	0.231	0,287	1.000
p(one tailed)	0.330	0.150	0.098	-
N	22	22	22	22

The results of this correlation show that there is a significant positive correlation between the attitude of paper-based participants and their essay scores (rho=0.42, p=03). This is not seen in the online participants where there is no significant correlation (rho==0.099, p=0.33). In order to investigate this finding further, an analysis of any difference in the attitude of those learners with mean high and low scores in their essays for both groups.

Table 12 below shows the mean rankings for the attitudes of learners classified as high and low achievers based on their essay scores, divided at the midpoint.

Table 12. Mean Ranking of the Attitude of Learners Classified as High and Low Performers

Tutor Marked Lower	11	15.14
Tutor Marked Upper	11	21.05
Online Marked Lower	11	26.00
Online Marked Upper	11	27.82
Total	44	-

A Kruskal Wallace test was performed on the data summarized above. The result of this analysis indicated that there was a significant difference between the attitude of the four groups (Chi-Square 6.497, df=3, p one-tailed=0.05). Post hoc analysis was performed using Mann Whitney U tests to test

the signicance of the mean rankings between the individual groups. The results of this analysis are presented in Table 13 below.

Table 13. Summary of the Post hoc Tests Carried out on the Data Summarized in Table 12 above

Group	Condition	N	Mean Rank	Mann-Whitney U	P (2 tailed)	
Tutor Based	Tutor Low	11	9.82	42.0	0.22	Not
Low v High	Tutor High	11	13.18			significant
Online Based	Online Low	11	10.73	52.0	0.58	Not
Low v High	Online High	11	12.27			significant
Online High	Paper Low	11	13.27	41.0	0.20	Not
v Paper Low	Online High	11	9.73			significant
Online Low	Online Low	11	12.86	45.0	0.32	Not
v Tutor High	Tutor High	11	10.14			significant
Online Low	Online Low	11	14.41	28.5	0.04	Significant
v Tutor Low	Tutor Low	11	8.59			

The results of this analysis show that there was a significant difference between online automated and tutor marked groups for those classified as low achievers. There was no other significant difference. Low achievers following the online system rated it higher than the tutor marked group. This may be due to several factors including the feedback provided by the system. Feedback and reflection as well as a summary of the results are discussed in the next section.

5. Discussion and Conclusion

In order to integrate automated essay marking into a Blended Learning context, it is important to show that it is able to perform at least as well as traditional methods. It must be as fair as traditional methods, not disadvantaging students. It should mark accurately when compared to essays marked by tutors. It should provide useful feedback that compares well to that provided by tutors, leading to improvement in performance over the duration of the course. The attitude of learners to the system should be at least as good as that of learners to the traditional tutor marked system. Tutors and learners should have confidence in the system. This is especially true of tutors if it is to be integrated successfully in a Blended Learning context. Our research has shown that the automated system marks accurately and fairly and that learners improve their performance over the duration of the course. Their attitude to the automated system was measured and shown to be comparable or better than the attitude of the control group to the tutor marked system.

Both groups were required to undertake reflective activities such as reflecting on their individual feedback and evaluation of their writing. The AES system provided greater opportunity for this. The AES system provided immediate feedback as soon as the students submitted their work. It allowed adequate time for the students to do as many corrections as they thought necessary. The system could then provide continuous suggestions to improve their work. In contrast, the traditional approach was time-consuming and required teachers to spend a lot of time and effort. Feedback and evaluation in the AES system was quantitatively different from that provided by the tutor. The fact that the AES system performed similarly or better than the traditional system in terms of scores obtained and attitude suggests that this feedback and reflective process was effective.

It may be argued that the significant difference in performance is due to the automated system marking "softer" than the tutor system. This indeed may be the case. It was also noted that the control group had a slightly lower pre-test mean score than the experimental group (although not significant). Future research is planned that will look to investigate these issues with larger groups that will be better matched and have less variance. The attitude of tutors to the system will be explored as this factor is essential in the implementation of the basic level blend.

References

Barker, T. (2011). An Automated Individual Feedback and Marking System. *Electronic Journal of E-Learning*, *9*, 1-14.

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

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, *18*(1), 7-24. https://doi.org/10.1016/j.asw.2012.10.002

Dikli, S., & 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

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

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

Joundy Hazar, M., Hussein Toman, Z., & Hussein Toman, S. (2019). Automated scoring for essay questions in E-learning. *Journal of Physics: Conference Series*. https://doi.org/10.1088/1742-6596/1294/4/042014

Kember, D., McNaught, C., Chong, F. C. Y., Lam, P., & Cheng, K. F. (2010). Understanding the ways in which design features of educational websites impact upon student learning outcomes in BL environments. *Computers and Education*, *55*(3), 1183-1192. https://doi.org/10.1016/j.compedu.2010.05.015

Lim, D. H., & Morris, M. L. (2009). Learner and instructional factors influencing learning outcomes within a blended learning environment. *Journal of Educational Technology & Society*, *12*(4), 282-293.

Manzanares, M. C. S., Sánchez, R. M., García Osorio, C. I., & Díez-Pastor, J. F. (2017). How do B-learning and learning patterns influence learning outcomes? *Frontiers in Psychology*, *8*(MAY), 1-13. https://doi.org/10.3389/fpsyg.2017.00745

Owston, R., York, D., & Murtha, S. (2013). Student perceptions and achievement in a university blended learning strategic initiative. *The Internet and Higher Education*, *18*, 38-46. https://doi.org/10.1016/j.iheduc.2012.12.003

Qian, L., Zhao, Y., & Cheng, Y. (2019). Evaluating China's automated essay scoring system iWrite. *Journal of Educational Computing Research*. https://doi.org/10.1177/0735633119881472

Raczynski, K., & Cohen, A. (2018). Appraising the scoring performance of automated essay scoring systems—Some additional considerations: Which essays? Which human raters? Which scores? *Applied Measurement in Education*, *31*(3), 233-240. https://doi.org/10.1080/08957347.2018.1464449

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

Ramineni, C., & Williamson, D. M. (2013). Automated essay scoring: Psychometric guidelines and practices. *Assessing Writing*, *18*(1), 25-39. https://doi.org/10.1016/j.asw.2012.10.004

Rupp, A. A., Casabianca, J. M., Krüger, M., Keller, S., & Köller, O. (2019). Automated essay scoring at scale: A case study in switzerland and germany. *ETS Research Report Series*, *2019*(1), 1-23. https://doi.org/10.1002/ets2.12249

Shanghai Jian Qiao University. (2019). *Blended Learning*. Retrieved December 12, 2019, from https://en.gench.edu.cn

Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing*, *20*, 53-76. https://doi.org/10.1016/j.asw.2013.04.001

Sun, F. (2014). The Application of Schema Theory in Teaching College English Writing. *Theory and Practice in Language Studies*, *4*(7), 1476-1482. https://doi.org/10.4304/tpls.4.7.1476-1482

Toranj, S., & Ansari, D. N. (2012). Automated Versus Human Essay Scoring: A Comparative Study. *Theory and Practice in Language Studies*, *2*(4), 719-725. https://doi.org/10.4304/tpls.2.4.719-725

Wang, E. L., Matsumura, L. C., Correnti, R., Litman, D., Zhang, H., Howe, E., … Quintana, R. (2020). eRevis(ing): Students' revision of text evidence use in an automated writing evaluation system. *Assessing Writing*, 100449. https://doi.org/10.1016/j.asw.2020.100449

Weigle, S. C. (2013). English language learners and automated scoring of essays: Critical considerations. *Assessing Writing*, *18*(1), 85-99. https://doi.org/10.1016/j.asw.2012.10.006

Westera, W., Dascalu, M., Kurvers, H., Ruseti, S., & Trausan-Matu, S. (2018). Automated essay scoring in applied games: Reducing the teacher bandwidth problem in online training. *Computers and Education*, *123*(April), 212-224. https://doi.org/10.1016/j.compedu.2018.05.010

Wilson, J., & Roscoe, R. D. (2020, 2019). Automated writing evaluation and feedback: Multiple metrics of efficacy. *Journal of Educational Computing Research*, *58*(1), 87-125. https://doi.org/10.1177/0735633119830764

Yang, Y. F. (2010). Students' reflection on online self-correction and peer review to improve writing. *Computers and Education*, *55*(3), 1202-1210. https://doi.org/10.1016/j.compedu.2010.05.017