2012 International Conference on Language, Medias and Culture
IPEDR vol.33 (2012) © (2012) IACSIT Press, Singapore

Automated Versus Human Essay Scoring: A Comparative Study

Somaye Toranj[1+], Dariush Nejad Ansari [2], Omid Tabatabaei[3]

[1]Islamic Azad University, Najafabad Branch, Iran
[2] University of Isfahan, Iran
[3]Islamic Azad University, Najafabad Branch, Iran

Abstract. The study evaluated two methods of essay scoring. Automated essay scoring (AES) system and human scoring. About 60 Iranian intermediate EFL learners were selected on a Standard English proficiency test (Allen 2004). Afterwards, they were randomly assigned to two groups of 30, experimental and control group. Participants in experimental group received the AES scoring, and control group, received the human scoring. Statistical analyses of the results revealed that 1) AES tool results in significant improvement of L2 learners writing achievement, 2) Results from questionnaire show that Students ware favor about using AES tool Hence, the findings of this study indicate that using AES tools can help teachers ease their big teaching students to improve their writing and it can be used as an educational tool on classrooms.

Keywords: Automate essay scoring (AES), Human scoring, Correlation, Writing improvement

1. Introduction

Writing is one of the most important skills that students need to develop, and the ability to teach writing is central to the proficiency of a well-trained language teacher [10]. New technologies have played an important role in the teaching of writing; writing teachers are often faced with these technologies. Using technology can change student writing behaviors. Writing assessment and providing feedback to students is often seen as one of the teachers' most important tasks.

Computerized feedback has been researched in studies as an alternative for enhancing the effectiveness of feedback. Researchers have found problems with the quality of feedback given by teachers; because of lack of time and large classes, teachers sometimes fail to give timely and precise feedback. In spite of the ample positive effects of feedback, these issues can critically and seriously limit the benefits of feedback. Understanding this problem, researchers and educators began to pay serious attention to the automated essay scoring system because of its potential as a mechanism for consistent and prompt feedback and essay grading.

Automated essay scoring (AES) is the ability of computer technology to evaluate and score written prose [14]. With the advent of new technologies, AES systems were developed to assist teachers' classroom assessment and to help overcome time, cost, reliability, and generalizability issues in writing assessment.

The research on AES has revealed that computers have the capacity to function as a more effective cognitive tool [3]. However, responding to student papers can be a burden for teachers. Particularly if they have large number of students and if they assign frequent writing assignments, providing individual feedback to student essays might be quite time consuming. AES systems can be very useful because they can provide the student with a score as well as feedback within seconds [12]. Previous research studies have demonstrated that a high score agreement rate could be achieved between human raters and automated scoring systems [11], [4]. The present study examined the relationship between AES and human scoring in order to determine usefulness of AES for writing assessment.

2. Review of Literature

[+] Somaye Toranj Tel.: + 980374906863; fax: +983812227301.
E-mail address: m2ehdi@hotmail.com

Research in the field of automated essay scoring began in the early 1960s. [13]. Burstein states that Educational Testing Service (ETS) has been conducting research in writing assessment since 1947. ETS administered the Naval Academy English Examination and the Foreign Service Examination as early as 1948 (Educational Testing Service, 1949-1950), and the Advanced Placement (AP) essay exam was administered in the spring of 1956 [15].

Attali, Bridgeman and, Trapani (2010) stated that essay writing assessments are often favoured over measures that assess student's knowledge of writing conventions, because they require students to produce a sample of writing and as such are more "direct." However, a drawback of essay writing assessments is that their evaluation requires a significant and time-consuming effort. These difficulties have led to a growing interest in the application of automated natural language processing techniques for the development of automated essay scoring (AES) as an alternative to human scoring of essays. Even basic computer functions, i.e. word processing, have been of great assistance to writers in modifying their essays [2].

AES affords the possibility of finer control in measuring the writing construct [5], [6]. The research on AES has revealed that computers have the capacity to function as a more effective cognitive tool (Attali, 2004) [3]. There are several different types of AES systems widely used by testing companies, universities, and public schools, Dikli (2006) discussed the following systems: Project Essay Grader™ (PEG), Intelligent Essay Assessor™ (IEA), CriterionSM, e-rater®, IntelliMetric™, MY Access® and BETSY [7].

According to Shermis, Burstein, Higgins and Zechner (2010), three major automated essay scoring were developed. The Educational Testing Service (ETS) has *e-rater*® which is a component of *Criterion*^{SM}, Vantage Learning has developed *Intellimetric*™ which is also part of an electronic portfolio administration system called *MyAccess!*^{TM} and Finally, Pearson Knowledge Technologies supports the *Intelligent Essay Assessor*™ which is used by a variety of proprietary electronic portfolio systems. As they stated that "all AES engines have obtained exact agreements with humans as high as the mid-80 and adjacent agreements in the mid-high 90's--slightly higher than the agreement coefficients for trained human raters" [15].

AES offers many advantages as increasing scoring consistency, introducing varied high-stakes assessments, reducing processing time and keeping the meaning of "Standardization" by applying the same criteria to all the answers. These systems have some disadvantages like extracting variables that are not important in evaluation operation, the lack of personal relationship between students and evaluators and the need for a large corpus of sample text to train the AES model [9]. According to Bennet and Ben-Simon (2005), "automated essay scoring has the potential to reduce processing cost, speed up the reporting of results, and improve the consistency of grading". As they stated the National Commission has recognized the potential value of this technology on Writing in Americas schools and colleges, with recommends research and development of AES system for standardized tests [5].

This study, therefore, sought answer to the following questions:

1. Does automated scoring using AES tool result in significant improvement of L2 learners writing achievement?
2. What are learner's attitudes toward AES tool?

3. Methodology

3.1 Participants

The participants of this study, selected through random sampling, consisted of 60 intermediate EFL learners majoring in English teaching at Shahrekord Azad University. The participants were classified into two groups after administrating the Oxford Placement Test (OPT): group one as the experimental and the other one as the control group. The experimental group, including 30 participants, received electronic scoring and the control group, including 30 participants, received human scoring.

3.2 Materials

In this study, four types of materials were used:

1- The Oxford Placement Test (OPT) developed by Allen (2004),

2- The Electronic writing rate Whitesmoke^{TM} which is one of the qualities writing enhancement software that using Artificial intelligence technology (AI) Natural Language Processing (NLP),

3- Writing quality assessment checklist, and

4- Questionnaire for learner interviews

In this study, one questionnaire was administered by researcher to indicate students' attitudes towards receiving AES tool. A group of students that received AES tool participated in this part. The questionnaires were ten questions consisted of two open-ended and eight Likert-scaled questions.

3.3. Procedure

This study was conducted with 60 intermediate EFL learners in Shahrekord Azad University. To collect the data, first, a multiple-choice proficiency test, (i.e., OPT developed by Allen, 2004) was administered. According to the scoring guidelines by Allen (2004), those whose scores in the test were among 60-75 were considered as the intermediate-level participants of this study [1]. The results showed the homogeneity of the juniors who were classified into two groups: Experimental Group and Control Group. The experimental groups (n = 3o) received the AES scoring, the control group (n = 30), received the human scoring. After receiving OPT exam both group have pretest - They were given a pretest in order to check their writing homogeneity. It was an essay writing task scored by two experienced teachers - , three writing tasks and a posttest all in five sessions, both group had the five topics same as writing prompts in the same genre. The time for writing an essay was 45-60 minutes. In the last session, the juniors on experimental group had a questionnaire to express their feeling and experiences with AES tool. The questionnaires had 2 open- ended and eight likret- scale questions. They had 5-10 minutes to complete the questionnaire.

4. Results

4.1. First research question:

As seen in Table 1 the two-tailed P value is less than 0.0001 by conventional criteria, this difference is considered to be extremely statistically significant. We can see the improvement of AES group scores in contrast of human scoring group. The AES evaluation such as teacher behavior and it concluded that AES can affect students' writing.

Table 1 Independent Samples t Test for the Juniors posttest

	Levene's Test for Equality of Variances		t-test for Equality of Means						
								95% Confidence Interval of the Difference	
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Equal variances assumed	15.548	.000	5.6591	58	0.0001	-12.0000	2.120	-16.2446	-7.7554
Equal variances not assumed			5.6591	49.1667	0.0001	-12.0000	2.120	-16.2446	-7.7554

Table 2 Group Statistics of AES and Human rater posttest

Group	N	Mean	Std. Deviation	Std. Error
Posttest AES	30	61.17	7.00	1.28
Human rater	30	49.1667	9.2656	1.6917

4.2. Second research question:

One sample T-Tests was used for analysis of the questionnaire; it compared the means of the two groups. As can be seen in the following tables students' were ensured about software scores. In general, the mean of all questions can be concluded that students were favor to use AES tools.

Table 3 One-Sample Test

	Test Value = 3			
	T	**df**	**Sig. (2-tailed)**	**Mean Difference**
S1	-1.242	29	.224	-.33333
S2	2.262	29	.031	.40000
S3	.724	29	.475	.13333
S4	1.795	29	.083	.30000
S5	.000	29	1.000	.00000
S6	1.980	29	.057	.33333
S7	1.490	29	.147	.26667
S8	.611	29	.546	.13333

Table 4 One-Sample Statistics

	N	**Mean**	**Std. Deviation**	**Std. Error Mean**
	30	3.1542	.71142	.12989

Table 5 One-Sample Test

	Test Value = 3			
	t	**Df**	**Sig. (2-tailed)**	**Mean Difference**
	1.187	29	.245	.15417

5. Discussion and Conclusion

To address the first research question, the following null hypothesis was tested:

H_{01}: AES does not result in significant improvement of learner's writing achievement.

The first null hypothesis was rejected. The results showed the effect of AES on improvement students essay writing (t =1.017, df = 58, P <0.05). The AES evaluation behaved such as teacher, and it concluded that AES can affect students' writing. It was revealed that the use of AES benefited the students in writing essays. In this stage the interreliability of teacher scoring was as independent factor and essay writings were as dependent factor. The results indicated that AES could score like teachers and could use in class for assessment essay writing. A comparison of mean scores of posttests in AES group and human scoring group, displays that the mean scores of the AES group has an increase of scores in contrast of human scoring group posttest(AES posttest : 61.17, human scoring: 49.1667) . It revealed that the students in the AES group wrote better essays in comparison with the students in human scoring group and it showed the role of AES to improvement of writing achievement.

Another research question was that what are learner's attitudes toward AES?

Most prior AES research fold into two categories -- the technical features of natural language processing (NLP), and reliability studies based on comparing human graders with an AES program rather than on students' attitudes and responses to and experiences with automated essay scorers. As mentioned before the questionnaire was used to look at the basic attitudes and opinions of the group of participant that scored their

essay by AES tool, this questionnaire could be helpful for future studies. Structured questionnaire just for AES group, it was associated with the context of their essay writing class that use AES tool as scorer. The main reason of collect questionnaire was to analyze the feedback of a new method of scoring in Iran, IELTS and TOFFLE institutes' uses computer to score, but using AES tool for improving writing is not common in Ministry of Education, and educational organization. Researcher must found feedback of this new method for further programming.

As seen in results, analyzing the questionnaire indicated that students were ensured about AES values and feedbacks, and they were like to accept this method. In summary analyzes questionnaire showed, the valuable benefits of AES in the classroom include increased motivation for students and easier classroom management for teachers.

The finding of the present study can lead to several important conclusions. The most important one is that computers could be useful in helping teachers ease their big teaching loads in some way and this method could help students to improve their writing. This study also has helped, to believe that automated essay scoring could be use as an education tools in classes.

Based on the results of the study, writing teachers need to be equipped with more recent developments in the field of e- rating. They should be aware of new methods of teaching writing to lead student to creativity in writing. Good writing skills are increasingly seen as vital to equip learners for success in this century.

6. References

[1] Allen, Dave. (2004). *The Oxford Placement Test*. Oxford: Oxford University Press.

[2] Y., Bridgeman, B., and, Trapani, C. (2010). Performance of a Generic Approach in Automated Essay Scoring. *Journal of Technology, Learning, and Assessment, 10*(3).

[3] Attali, Y. (2004). Exploring the feedback and revision features of Criterion. *National Council on Measurement in Education (NCME),* San Diego, CA.

[4] Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater v.2. *Journal of Technology, Learning, and Assessment 4 (3)*.

[5] Bennett, R. E. (2004). *Moving the field forward: Some thoughts on validity and automated scoring* (ETS Research Memorandum No. RM-04-01). Princeton, NJ: ETS.

[6] Bennett, R. E., & Ben-Simon, A. (2005). *Toward theoretically meaningful automated essay scoring.* Unpublished final project report, Educational Testing Service.

[7] Dikli, S. (2006). An overview of automated scoring of essays. *Journal of technology, learning, and assessment, 5*(1). Retrieved July 15, 2007, from http://www.jtla.org.

[8] Elliot, S. (2003). IntelliMetric™: From here to validity. In M.D. Shermis & J.C. Burstein (Eds.), *Automatic essay scoring: A cross-disciplinary perspective* (pp. 71-86). Mahwah, NJ: Lawrence Erlaum Associates.

[9] Hamp-Lyons, L. (2006). Feedback in portfolio based writing courses. In K. Hyland & F. Hyland (Eds.).

[10] Hyland, K. (2003). *Second language writing*. New York: Cambridge University Press.

[11] Kukich, K. (2000). Beyond automated essay scoring. *IEEE intelligent systems, 15*(5), 22-27.

[12] Page, E.B. (1966). The imminence of grading essays by computers. *Phi Delta Kappan, 47*, 238-243.

[13] Page, E. (1994). *Computer Grading of Student Prose, Using Modern Concepts and Software* Journal of Experimental Education, 62(2), 127-142.

[14] Shermis, M., & Burstein, J. (Eds.) (2003). *Automated essay scoring: A cross-disciplinary perspective*.Hillsdale, NJ: Lawrence Erlbaum.

[15] Shermis, M., & Burstein, J., & Higgins, D., and Zechner, K. (2010). Automated Essay Scoring: Writing Assessment and Instruction.