ISSN 1799-2591
Theory and Practice in Language Studies, Vol. 2, No. 4, pp. 719-725, April 2012
© 2012 ACADEMY PUBLISHER Manufactured in Finland.
doi:10.4304/tpls.2.4.719-725

Automated Versus Human Essay Scoring: A Comparative Study

Somaye Toranj
Islamic Azad University of Njafabad Branch, Isfahan, Iran
Email: m2ehdi@hotmail.com

Dariush Nejad Ansari
University of Isfahan, Isfahan, Iran
Email: nejadansari@gmail.com

Abstract—**This study investigated the effects of automated essay scoring (AES) system on writing improvement of Iranian L2 learners. 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 reveal 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, 3) The results from the current study support the conclusion that the AES tool does not seem to correlate well with human raters in scoring essays. 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.**

Index Terms—**automate essay scoring (AES), human scoring, correlation, validity, reliability**

I. 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 (Hyland, 2003).

New technologies have played an important role in the teaching of writing; writing teachers are often faced with these technologies. Writing always needs some kinds of application of technology, whether pencil, typewriter, or printing press, and each innovation involves new skills applied in new ways (Lankshear & Snyder, 2000).

An effective teacher could make the best decisions about methods, materials, and procedures used in classroom. New technologies offer opportunities for learners to engage with the creative process of construction and for teachers to help them to make their writing processes more effective. Technology is not a method but a resource which can support a variety of approaches (Warschauer, 2002). Using technology can change student writing behaviors. According to Warschauer and Kern (2000), the use of computers in language teaching reflects a move from structural through cognitive to sociocognitive orientations to teaching.

Writing assessment and providing feedback to students is often seen as one of the teachers' most important tasks. Feedbacks allow students to see how others respond to their work and to learn from the responses. Assessment is not simply a matter of setting exams and giving grades. Scores and evaluative feedback contribute enormously to the development of an effective and responsive writing course. In recent years, computers have opened up new opportunities for providing feedback to writing and it offers teachers greater flexibility in education.

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 (Shermis & Burstein, 2006). 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 (Attali, 2004). 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 (Page, 2003). Previous research studies have demonstrated that a high score agreement rate could be achieved between human raters and automated scoring systems (Kukich 2000; Attali & Burstein 2006; Ben-Simon & Bennett 2007).

The present study examined the relationship between AES and human scoring in order to determine usefulness of AES for writing assessment.

II. Background to the Study

Automated essay scoring (AES) is defined as the computer technology that evaluates and scores written works (Shermis& Burstein 2003).

AES appears with different titles like automated essay evaluation, automated writing evaluation, automated essay grading, automated essay assessments and automated writing scoring.

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 (Hamp-Lyons 2001) and the need for a large corpus of sample text to train the AES model (Chung & O'Neil 1997).

Research in the field of automated essay scoring began in the early 1960s (Page, 1994). 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. Some of the earliest research in writing assessment laid the foundation for holistic scoring a scoring methodology used currently by ETS for large-scale writing assessments (see Coward, 1950 and Huddleston, 1952).

Attali, Bridgeman and, Trapani (2010) stated that essay writing assessments are often favored over measures that assess student̓s 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.

AES affords the possibility of finer control in measuring the writing construct (Bennett, 2004).

The research on AES has revealed that computers have the capacity to function as a more effective cognitive tool (Attali, 2004).

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.

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".

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 (National Commission on Writing, 2003, pp. 30-31).

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

1. Is there any correlation between scores assigned by AES tool and scores assigned by human raters?
2. What are learner's attitudes toward AES tool?
3. Does automated scoring using AES tool result in significant improvement of L2 learners writing achievement?

III. Methodology

A. 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.

B. Materials

In this study, four types of materials were used

1. The Oxford Placement Test (OPT) developed by Allen (2004). According to the scoring guidelines by Allen (2004), the scores among 60-75 were considered as the intermediate level.

2. The Electronic writing rater Whitesmoke^{TM} is one of the qualities writing enhancement software that using Artificial intelligence technology (AI) Natural Language Processing (NLP); it can correct errors that commonly occur in the natural flow of writing.. Whitesmoke^{TM} has some feature as below:

- Comprehensive checkers for grammar, spelling, punctuation, structure and style
- Translation of full texts
- Multi-lingual dictionary
- Artificial intelligence technology(AI)
- Templates on various writing styles for business letters, resumes, finance writing, greetings, etc.

3. Writing quality assessment checklist

In this study to reduce scorer errors and attend reliability and validity of the scores given to each paper, for scoring the students' papers**,** Roebuck's Analytic scoring Rubrics, modified by Maftoon & Rabiee (2006) as a writing assessment check list were used.

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.

C. 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.

IV. Results

1. First research question: Pearson correlation coefficient test was selected for determining the correlations between AES tools and human scoring. The results of the correlational analyses indicated that there was no statistically significant correlation between Whitesmoke™ and human scoring (*r* = 0.121). As seen in Table 4.6 there is no relationship between AES tool and human rater.

TABLE-1
THE RESULTS OF CORRELATIONAL MEASUREMENT

Result	N	sig	r
	30	0.522	0.121

2. Second research question: One sample T-Tests was used for analysis of the questionnaire; it compared the means between 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 -2
ONE-SAMPLE STATISTICS

	N	Mean	Std. Deviation	Std. Error Mean
S1	30	2.6667	1.47001	.26839
S2	30	3.4000	.96847	.17682
S3	30	3.1333	1.00801	.18404
S4	30	3.3000	.91539	.16713
S5	30	3.0000	.98261	.17940
S6	30	3.3333	.92227	.16838
S7	30	3.2667	.98027	.17897
S8	30	3.1333	1.19578	.21832

Table -3
One-Sample Statistics

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

Table – 4
One-Sample Test

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

Table-5
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

3. Third research question: As seen in Table 8 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-6
Group Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Experienced teacher	30	52.1917	11.50650	2.10079
AES tool	30	49.8000	5.79179	1.05743

Table-7
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

Table-8
Independent Samples t Test for the Juniors posttest

	Levene's Test for Equality of Variances		t-test for Equality of Means						
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the 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

V. Discussion and Conclusion

The purpose of this study was to consider the relationship between automated essay scoring (AES) and human scoring in order to determine the usefulness of AES for assessment writing tests in large community. This study was conducted with 60 intermediate EFL learners majoring in English teaching at Shahrekord Islamic Azad University in the second semester of the academic year 2011-2012.

A correlational research design was used to answer the first research question; correlations between AES performance and human raters' performance were examined. For answer to the second research question a questionnaire were prepared. Third research question were analyzed by using independent t test. This study produced a number of key findings in relation to the aim of the survey.

Results based on the correlational data analyses showed no statistically significant correlation between Whitesmoke^{TM} scoring and human scoring in terms of overall holistic scores. This finding did not corroborate most previous studies conducted AES systems, which reported strong correlations between AES and human scoring for overall ratings [14]. But it has some exceptions, like Wang and Brown studies (2007& 2008). In the next section the finding of the study correlated with each of the three research questions presented above.

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

H_{01}: There is no correlation between scores assigned by AES tool and scores assigned by human raters.

The first null hypothesis was not rejected. To evaluate this null hypothesis, Pearson correlation coefficient test was conducted.

Results based on the correlational data analyses showed no statistically significant correlation between Whitesmoke^{TM} scoring and human scoring in terms of overall holistic scores (n= 30, r = 0.121).

On the whole, the results from the current study support the conclusion that Whitesmoke^{TM} did not seem to correlate well with human raters in scoring essays these results may be attributed to the following factors:

1. Most of AES tools were designed for English language and its features. As mentioned on literature review, most studies conducted by AES tools were validation studies, which had split the pool of student writing samples from the same student population into two groups, using one group to build the scoring model and the other group as a validation data set. This means the scoring model built by the writing samples drawn from the same student population as the validation set might have extracted writing features idiosyncratic to the particular student population. Therefore, the scoring model could score the validation set with relatively high accuracy, but it is questionable whether its application is generalizable to other student populations who receive different writing instruction and have different writing experiences. It also true about Whitesmoke$^{TM;}$ AES tools didn't examined in Iran educational systems until now.

2. In the current study, Whitesmoke^{TM} scoring have not a significant correlation with human scoring, so it is possible that automated essay scoring tools tend to be more accurate in evaluating surface features of writing samples at the sentence level. In contrast, human raters are trained to regard surface features as one of the five dimensions of writing; they may not have weighted surface features as heavily as did AES tools. More importantly, human raters emphasize meaning-making and communicative contexts, which AES tools may still be incapable of identifying and evaluating. In many counties the interest in adopting AES tools increases, and as the development of AES technologies undergoes rapid changes, they still hold a promising future for writing assessment programs; therefore, continuous research and investigation in the validity and generalizability of the AES tools are inevitable. Also in Iran this studies is required.

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 a 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.

It should be indicated that during survey participants were enthusiasm about this method but using technology for them in class was a little strange and unusual. In summary analyzes questionnaire showed, the chief benefits of AES in the classroom include increased motivation for students and easier classroom management for teachers. The main point to bear in mind is that such automated systems do not replace good teaching but should instead be used to support it. This is particularly so with the instruction of weaker students, who may lack the requisite language and literacy skills to make effective use of automated feedback. To matching students with this method, at first they must learn to use technology as a way to improve their weakness' of learning, after that they found necessity of AES, they need to learn write for a variety of audiences, including not only computer scoring engines, but peers, teachers, and interlocutors outside the classroom, it could help them to improve writing.

The third research question was does automated scoring using AES tool result in significant improvement of learner's writing achievement?

H_{02}: AES does not result in significant improvement of learner's writing achievement.

The second 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.

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. The ability to communicate ideas and information effectively through the global digital network is crucially dependent on good writing skills.

ACKNOWLEDGMENT

Once more, we like to express our deepest gratitude to Dr. Omid Tabatabai of the English department at Najafabad Islamic Azad University for guiding us with his wisdom, expertise, and encouragement and Dr. Chaitanya Ramineni and F Jay Breyer Educational Testing Service team for their helps and insightful comments.

REFERENCES

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

[2] Attali, Y., Bridgeman, B., and, Trapani, C. (2010). Performance of a Generic Approach in Automated Essay Scoring. *Journal of Technology, Learning, and Assessment*, 10.3, 4-7.

[3] Attali, Y. (2004). Exploring the feedback and revision features of Criterion. *Journal of Second Language Writing, 14.3, 191-205.*

[4] Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater v.2. *Journal of Technology, Learning, and Assessment* 4.3, 13-18.

[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. *Journal of Technology, Learning, and Assessment* 6.1, 1-47.

[7] Chung and O'Neil, (1997). Methodological Approaches to Online Scoring of Essays CSE Technical report 461, National Center for Research on Evaluation, Standards, and Student Testing.

[8] Dikli, S. (2006). An overview of automated scoring of essays. *Journal of technology, learning, and assessment, 5*.1, 5-22.

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

[10] Hamp-Lyons, L. (2006). Feedback in portfolio based writing courses. In K. Hyland & F. Hyland (Eds.), *Feedback in second language writing: Contexts and issues*. Cambridge: Cambridge University Press, 140-161.

[11] Hyland, K. (2003). Second language writing. New York: Cambridge University Press. P. xv-29.

[12] Kukich, K. (2000). Beyond automated essay scoring. *IEEE intelligent systems,* 15.5, 22-27.

[13] Landauer, T.K., & Laham, D. (2000). The Intelligent Essay Assessor. *IEEE intelligent systems*, 15.5, 27-31.

[14] Page, E. B. (2003). Project Essay Grade: PEG. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Lawrence Erlbaum Associates, 43–54.

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

[16] Page, E. (1994). Computer Grading of Student Prose, Using Modern Concepts and Software *Journal of Experimental Education*, 62.2, 127-142.

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

[18] Shermis, M., & Burstein, J., & Higgins, D., and Zechner, K. (2010). Automated Essay Scoring: Writing Assessment and Instruction, 6-12.

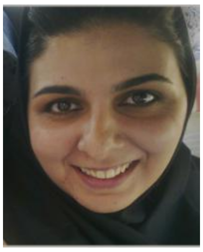
[19] Wang, J., & Brown, M. S. (2008). Automated essay scoring versus human scoring: A correlational study. *Contemporary Issues in Technology and Teacher Education*, *8*.4, 310-325.

[20] Wang, J. & Brown, M. (2007). Automated essay scoring versus human scoring: A comparative study. *Journal of Technology, Learning, and Assessment*, 6.2, 20-22.

[21] Warschauer, M. (2002). A review of *Language and the Internet* by David Crystal. *Education, Communication, and Information*, 2.2, 241-244.

[22] Warschauer, M., & Kern, R. (Eds.). (2000). Network-based language teaching: Concepts and practice. Cambridge: Cambridge University Press, 41-58.

[23] Wresch, W. (1993). The imminence of grading essays by computers - 25 years later. *Computers and composition* 10.2, 45-58.



Somaye Toranj Place of Birth: Shahrekord, Iran, Date of Birth: 8/1/1981, Gender: Female, Marital Status: Single

She is originally from Shahrekord, Iran. She holds a M.A. degree in TEFL in 2010 from Islamic Azad University, Najafabad Branch.

She is currently a full-time high school teacher of English in Iran, and teaches undergraduate (BA) English majors at the local Payamenoor University. Her research interests are in the application of second language acquisition theories to the teaching of foreign languages, language acquisition, testing and assessment, teaching skills, technology use in the foreign language classroom.

Dariush Nejad Ansari The assistant professor Dariush Nejad Ansari is an academic member of the University of Isfahan. He has been teaching at different levels in English Department of the Faculty of Foreign Languages. He graduated with an MA in applied linguistics from Tarbiat Modarres University in **1996** and completed his PhD at Allame Tabatabaei University in TOEFL in **2009**. His areas of interest are issues in language teaching/learning and translation.