European Journal of English Language Teaching
ISSN: 2501-7136
ISSN-L: 2501-7136
Available on-line at: www.oapub.org/edu

doi: 10.5281/zenodo.3461784 Volume 5 | Issue 1 | 2019

WEB 2.0 AUTOMATED ESSAY SCORING APPLICATION AND HUMAN ESL ESSAY ASSESSMENT: A COMPARISON STUDY

Mohammad Radzi Manap[1],
Nor Fazlin Ramli,
Aini Akmar Mohd Kassim
PhD, Akademi Pengajian Bahasa,
Universiti Teknologi MARA,
Shah Alam, Selangor,
Malaysia

Abstract:
Educators from all levels of education have comfortably embraced ICT for decades. More and more applications are being developed which has direct and indirect advantages on learning and teaching. What is more interesting is the fact that more free applications are now being made available to be used by the public. In this paper, the focus is on the comparison between a free Automated Essay Scoring (AES) web-based application called 'PaperRater.com' and human assessment by English instructors from a public university in Malaysia. Ten selected human assessed essays were assessed by individual lecturer (IL) and also a group of lecturers (GL). Those essays were then fed into 'PaperRater.com' and a comparison was made in terms of total scores generated by the application and the ones by the instructors. A descriptive statistical analysis was carried out to compare the scores. As a result, assessment for both types of essay recorded a difference of between 0.3 to 38.7 marks where bigger disparity was recorded between human assessed and computer based (AES) assessment. Overall, however, the strength of the relationship between GL and PR, based on the Pearson's product moment correlation coefficient, was recorded at 0.678407 which means that there is a moderate positive linear relationship.

Keywords: Computer Assisted Language Learning, e-learning, web-based 2.0 application, Automated Essay Scoring

1. Introduction

Free applications in the form of web-based, web-based mobile or native applications specifically for language learning are common nowadays. Their existence has benefitted

[1] Correspondence: email mradzim7@gmail.com, moham830@uitm.edu.my

many language instructors as teaching tools in their language classes. Not only that the use of those applications is free, but they also are able to assist language teachers to vary their approach in teaching as well as reducing their burden dealing with many academic and administrative works. Most importantly, the involvement of the use of ICT in the teaching and learning process has always resulted in the increase of motivation and interest in learning among language learners. This paper will investigate the implications in terms of suitability and usefulness of using one of the free applications available for essay checker among the students and English teachers in a public university in Malaysia and the automated essay scoring AES used is called 'PaperRater.com'. It is one of many applications categorized as AES or automated writing evaluation (AWE) available nowadays.

Attali and Burstein (2006) reported that AES has become a viable and reliable alternative complementing human scoring since as early as 1966 through the works by Page (1966), Project Essay Grade (Page, 1994), e-rater (Burstein et al., 1998), Intelligent Essay Assessor (Landauer, Foltz, & Laham, 1998) and Intellimetric (Elliot, 2001). In 1973, according to Shermis (2010), a successful AES system was programmed which required punch cards and mainframe computer. The advancement of technology has made the capability of such systems to be even closer to human capabilities in assessing essay or writing. In describing AES, Dikli (2006) has identified 4 types of commonly used AES systems that are Essay Grade (PEG), Intelligent Essay Assessor (IEA), E-rater and Intellimetric. Out of these four types, the Essay Grade (PEG) is known to be the first AES system built in AES history while Intellimetric is the first AES system that is constructed based on artificial intelligence.

In this study, a comparative investigation was carried out in order to reveal the differences in terms of total score awarded by 'PaperRater.com' and English lecturers in this particular institution of higher learning. This paper attempts to answer these research questions:

1) Is there a difference in total scores of essays marked by individual lecturers (IL), group lecturers (GL) and paperrater (PR)?
2) Is there a correlation between the scores of essays marked by individual lecturers (IL), group of lecturers (GL) and paperrater (PR)?
3) To what extent can paperrater be used as a viable assessment tool?

1.1 Limitations

There are several limitations to this study. Firstly, there is a limited number of essays utilized in the analysis (10 essays). However, the results from the statistical analysis should not be dismissed. This study focuses on the usability of an automated essay rater which lends to a detailed exploration of the total scores. The number of essays itself was determined and handpicked by the human raters themselves as a benchmark or reference before they mark the rest of the essays. In addition, 'PaperRater.com' capability is not fully tested as the one used in this study is not the premium version

which has extra features which could make the difference. Further studies on later versions of 'PaperRater.com' are recommended.

2. Literature Review: What Is 'PaperRater.com'?

'PaperRater.com' is a free web-based application used for assessing written materials of different sorts. It is used as a writing tool which is powered by natural language processing (NLP), artificial intelligence (AI), machine learning, information retrieval (IR), computational linguistics, data mining, and advanced pattern matching (APM). 'PaperRater.com' is one of many Web 2.0 applications that could be useful to English instructors from various levels of education in assisting them to assess writing works of their students. 'PaperRater.com' offers free services that include Plagiarism Detection, Auto Grader, Spelling and Grammar Check, Style and Word Choice Analysis, Readability Statistics, Title Validation and Vocabulary Builder tool. This type of AES is also known as AWE or Automated Writing Evaluation. Using the term CBEM or Computer Based Essay Marking, Saadiyah (2003) also listed down a number of other systems for example, Methodical Assessment of Reports by Computer (Marshall and Baron 1987), Markin 32 (Holmes 1996), Project Essay Grader (Page, Fisher and Fisher 1968; Hller 1998; Page Petersen 1995; Shemis et al.), Intelligent Essay Assessor (Landauer, Foltz and Laham 1998), and Criterion (Burstein, Chodorow and Leacock 2003). These are among the available AES in the market. It depends on the users' preference and needs on which application to use.

2.1 Why 'PaperRater.com'?

Generally, there is almost no free AES in the open market that can cater to the specific assessment required by the users. In previous researches, there have been attempts in customizing applications that fulfill the requirements and needs of the instructors. However, some of the attempts failed due to the insufficient amount of essay to feed the database system that resulted the scoring to be inaccurate (Uzun, 2018). In some other cases as reported by Saadiyah Darus et al. (2000), through the survey conducted with 6 essay marking systems has concluded that most of these systems are promising but further research needs to be carried out in areas related to theory and practice, assessment, pedagogies and their suitability with the writing needs of the learners of different racial and cultural background including the writing requirements of the institutions they belong to.

In reality, when it comes to the actual use of AES, instructors as well as learners are in the look for the ones that are free and ready to be used. Apart from the issues discussed above, affordability has become an issue when learners are required to subscribe to the application and the hassle for the instructors to make sure total involvement among the learners due to financial issues. However, at the same time, the advantages or using AES in classroom is a boon to many as they are able to solve multiple issues faced by instructors in providing prompt feedback that are crucial to

learners as a motivation boost. Due to these facts, 'PaperRater.com' is chosen to be highlighted in the research as it represents the freely available AES to be used by the vast majority.

Another crucial element that is considered in using 'PaperRater.com' as the chosen AES in this study is simply because 'PaperRater.com' provides total score for each essay in their reporting, an aspect of reporting that is not available in most free AES. This criteria is seen crucial and important to the learners as a quick check on their progress in writing. At the same time the scores generated by papaerrater.com allows instructors and researchers to make direct comparison to human raters.

2.2 Previous Study on Comparison between Human and Computer Essay Assessment

Looking into the previous works on comparing the human and automated essay scoring systems, this paper will list down 11 comparisons between human and machine assessments taken from various works from 2003 to 2014. The results are tabulated in Table 1.

Table 1: Comparisons between human and machine assessments

No	Machine Assessment	Reported by	Result
1.	Project Essay Grade (PEG)	Valenti, Neri, and Cucchiarelli (2003)	87 (correlation)
2.	Intelligent Essay Assessor (IEA)	Valenti, Neri, and Cucchiarelli (2003)	85-91 (agreem)
3.	Educational Testing service I	Valenti, Neri, and Cucchiarelli (2003)	93-96 (accuracy)
4.	Electronic Essay Rater (E-Rater)	Valenti, Neri, and Cucchiarelli (2003)	87-94 (agreem)
5.	C-Rater	Valenti, Neri, and Cucchiarelli (2003)	80 (agreem)
6.	BETSY	Valenti, Neri, and Cucchiarelli (2003)	80 (accuracy)
7.	Intelligent Essay Marking System	Valenti, Neri, and Cucchiarelli (2003)	80 (correlation)
8.	Automark	Valenti, Neri, and Cucchiarelli (2003)	93-96 (correlation)
9.	IntelliMetric™	Wang and Brown (2007)	non-significant mean score differences between AES and human scoring
10.	Whitesmoke	Toranj and Ansari (2012)	no significant correlation
11.	Criterion	Huang (2014)	weak correlation

2.3 AES and AWE Impacts on ESL

The use of automated essay scoring in the field of ESL has made its marks decades ago. In fact, the first automated essay scoring (AES) was developed by an English teacher,

Ellis Page, in 1966. Page called his invention as Page Essay Grade (PEG). In the beginning, PEG dealt with surface text features analysis like number of words, average sentence length until later it was able to include other more meaningful features like grammatical correctness and word choice. Such features are not only meaningful to human raters but they give great pedagogical impacts on the field of English language teaching and learning. It was found that major scoring engines roughly equivalent to human graders in reliability.

Things are getting brighter for English language instructors as more advanced software categorized as Automated Writing Evaluation (AWE). According to Chi-Fen Emily Chen and Wei-Yuan Eugene Cheng (2008), since the mid-1990s, the design of AWE programs has been improving rapidly due to the advancement of artificial intelligence technology, in particular natural language processing and intelligent language tutoring systems. Most importantly, these systems could help language instructors to reduce bottleneck in marking students' essays, provide immediate feedback to students, increase learner autonomy, support drilling and scaffold language learning. In a study by Grimes (2010), the essay scoring system has managed to increase the amount of writing by students to an average of 33% and Dikli (2006) found that the accuracy and reliability of the AES was proven to be high. Feedbacks from teachers proved automated writing evaluation managed to reduce their grading burden and supported individualized instruction (Warschauer & Grimes, 2008). AES has become the answer in overcoming some weaknesses among human raters especially in dealing with huge volume of essays. Apart from the need of human to be recruited, instructed on the use of the rubrics, certified of their rating competencies and closely monitored, human also tend to make mistakes and being bias. Mo Zhang (2013) summarizes the sources of human errors as listed in Table 2.

Table 2: Descriptions of Some Common sources of Human-Rater Errors and Biases

Severity/ Leniency	Refers to a phenomenon in which raters make judgments on a common dimension, but some raters tend to consistently give high scores (leniency) while other raters tend to consistently give low scores (severity), thereby introducing systematic biases.
Scale Shrinkage	Occurs when human raters don't use the extreme categories on a scale. Inconsistency Occurs when raters are either judging erratically, or along different dimensions, because of their different understandings and interpretations of the rubric.
Halo Effect	Occurs when the rater's impression from one characteristic of an essay is generalized to the essay as a whole. Stereotyping Refers to the predetermined impression that human raters may have formed about a particular group that can influence their judgment of individuals in that group.
Perception	Appears when immediately prior grading experiences. Difference influences a human rater's current grading judgments. RaterDrift Refers to the tendency for individual or groups of raters to apply inconsistent scoring criteria over time.

Source: Mo Zhang (2013).

Generally, consistency is the key factor of AES in assessing essays apart from other crucial aspects like accuracy and immediacy in getting feedback and results. However, on the other hand, there are also some drawbacks of AES for instance a study by Wang

and Brown (2007) revealed that the mean score by an AES called Intellimetric™ were found to be higher than the human raters'. Another study conducted by Saadiyah et al. (2003) has revealed that the subjects in her study found the feedback given by the system is useful and informative only to some extent. The feedbacks are also found not sufficient to help students to improve. Saadiyah (2003) suggested for a more customized system designed for Malaysian ESL learners.

In a more recent study conducted by Qui Yubing (2016), the automated essay scoring or rather referred to as Smart Essay Scoring (SES) called Pigai, claimed to be the biggest and probably the most influential SES system in China, has outlined some of the advantages as well as the disadvantages of Pigai. Apart from being able to offer immediate feedback, check plagiarism, ease the assessment process, and provide reliable grades, Pigai has found not to be able to provide the users the reasons for the sentence to be grammatically wrong. Furthermore, Pigai also has not been successful in checking sentence structure as well as not able to detect the coherence of an essay. Considering all these, Qui Yubing has concluded that language teacher should use Pigai wisely and need to explore to suitability of its use in English classrooms.

3. Methodology

This research adopts a quantitative study. In order to compare and describe the scores of AES scores and human rater scores, descriptive statistics are used. Descriptive statistics are applied to describe the fundamental features of the data in a study. They provide simple summaries about the sample and the measures. Together with simple graphics analysis, they can form a clear representation of the collected data (Creswell, 2009).

For this study, Pearson's correlation coefficient is used to measure the statistical relationship, or association, between the two continuous variables. In statistics, the correlation coefficient r measures the strength and direction of a linear relationship between two variables on a scatterplot. According to Moore, Notz, and Flinger (2013) and Hinkle et al. (2003) the value of r is always between +1 and –1. Table 3 illustrates the strength of association that guide the research based on the rule of thumb for interpreting the size of a correlation coefficient.

Table 3: Correlation Coefficient Table of Relationship

Size of correlation	**Interpretation**
.90 to 1.00 (-.90 to -1.00)	Very high positive (negative) correlation
.70 to .90 (-.70 to -.90)	High positive (negative) correlation
.50 to .70 (-.50 to -.70)	Moderate positive (negative) correlation
.30 to .50 (-.30 to -.50)	Low positive (negative) correlation
.00 to .30 (.00 to -.30)	Negligible correlation

3.1 Data Collection and Analysis

Data in this study were mainly gathered from students' essays. The essays selected in this study were chosen by the resource person of the particular course of which the paper is offered. Those chosen essays (10 essays) were then brought to a central marking committee to be assessed. This is done to ensure the validity of the scores. However, before the actual committee members sat for the meeting, copies of those selected essays were given to individual lecturers involved to be assessed individually based on the standard answer scheme given. On the day of the meeting, all members discussed various aspects of assessing the essays until they came to a consensus. Finally, final marks were given to all the essays. The task of the committee is to align the assessment of all its members/examiners so that a standard of marking or assessment is achieved. This is to ensure fairness, quality and standard of assessment to be adhered to by all involved. This is a normal assessment procedure practiced by this institution.

The ten essays were marked based on impression marking by 140 language instructors. Scores of the 10 essays were taken randomly from one individual instructor (one lecturer from the 140) to be later used as scores generated by individual lecturer (IL). Next, the average scores of the 10 essays generated from the assessment of the committee (average scores) were also recorded. In the analysis, these scores are labelled as scores from group of lecturers (GL). The third source of scores was later taken from scores generated by 'PaperRater.com'. In order to generate the scores, the 10 hand-written essays were transformed into Microsoft Word format through verbatim retyping and were run into 'PaperRater.com' to generate the scores. These scores represent score by the AES or the 'PaperRater.com' and labelled as (PR). In submitting those essays into 'PaperRater.com', the researcher had chosen 'college (undergraduate)' as the education level of the writer of the essays and chosen 'essay' as the type of paper to be submitted. Besides, the plagiarism detection was not selected for all the 10 essays. All this information is required by 'PaperRater.com' before any submission is to be made and fo any scores to be generated. The total scores of each essay by individual lecturers, group lecturers and paperrater are tabulated in Table 7.

3.2 Assessment Guides

All the ten essays were assessed holistically based on the rubrics of the Common European Framework (CEFR). Besides, 'PaperRater.com' also has its own different aspects of assessing the essay. Though the information on how exactly the essays were assessed by 'PaperRater.com' is not made known, the items in the report generated by 'PaperRater.com' are used to describe the aspects taken in for the assessment. Therefore, it is assumed that the assessment of the essays was based on the items generated in the report as provided in the tables below:

Table 4: Paperrater's Report Items

Paperrater's Report Items
Spelling
Grammar
Word choice
Style
Usage of Transitional phrases
Sentence Length Info
Sentence beginning
Readability Indices (Premium only)
Passive voice (Vocabulary words)
Grade (percentage)

The same set of essays is, however, assessed by the human raters based on the Common European Framework of Reference (CEFR) criteria which was adopted for the first time by the institution. The scoring based on CEFR is depicted in TABLE 5 and the recorded scores used in this analysis were based on the average scores given by the group of lecturers (GL) for each essay. This explains why the reported scores based on the CEFR to be in one decimal point.

Table 5: CEFR Grading guide used in a Standard English University Test

Score	**Level**	**User**
6	C2	Proficient User
5	C1	Proficient User
4	B2	Independent User
3	B1	Independent User
2	A2	Basic User
1	A1	Basic User

Besides the two assessment guides considered in this study, another grading guide that should be considered is the University's grading guide (TABLE 6). This will eventually become the most crucial guide as it is used to determine the actual scores of the students. The University's grading guide will be used as the final and tool to determine the disparity between all the assessors in this study.

Table 6: University's Grading Guide

90 — 100	A+	4.00	Pass
80 — 89	A	4.00	Pass
75 — 79	A-	3.67	Pass
70 — 74	B+	3.33	Pass
65 — 69	B	3.00	Pass
60 — 64	B-	2.67	Pass
55 — 59	C+	2.33	Pass
50 — 54	C	2.00	Pass
47 — 49	C-	1.67	Fail
44 — 46	D+	1.33	Fail
40 — 43	D	1.00	Fail

30 — 39	E	0.67	Fail
0 — 29	F	0.00	Fail

In order to have a meaningful interpretation of the comparison of scores from all the assessments (CEFR and 'PaperRater.com'), all scores will be transformed into percentage. Since the 'PaperRater.com' scores are already in the form of percentage, therefore, all the average ratings of essay using the CEFR will also be converted into percentage thus making the comparison to be more meaningful. Apart from this, reference will also be made on the university's grading status to determine range of acceptance. This means that any difference in terms of score can only be accepted if it does not affect the grade. For instance, if an essay is scored to have 80%, 84% and 89% by individual lecturer, central marking and 'PaperRater.com' respectively, the differences will be considered as acceptable since 80%, 84% and 89% are still within 'Grade A'. The benchmark of grading in this case will be based on the average scores by the lecturers.

4. Results

The findings in this study are divided into three sections:

1) The difference in total scores of essays marked by individual lecturers and group lecturers, and paperrater.
2) The correlation between the scores of essays marked by individual lecturers and group of lecturers and paperrater.
3) The extent paperrater can be used as a viable assessment tool.

4.1 The difference in total scores of essays marked by individual lecturers, group lecturers, and paperrater

The differences in total score of all the 10 essays marked by individual lecturer, group lecturer and 'PaperRater.com' are summarized in TABLE 7. Here, the differences are recorded in percentages.

Table 7: Difference in Essay Scores of All Three Raters

Essay	Individual Lecturer (IL) %	Group Lecturer (GL) %	'PaperRater.com' (PR) %	Difference (IL-GL) %	Difference (IL-PR) %	Difference (GL-PR) %
E1	4.0 /6 (66.7%)	4.4 /6 (73.3%)	80	**-6.6**	13.3	6.7
E2	3.2 /6 (53.3%)	4.6 /6 (76.6%)	82	**-23.3**	28.7	5.4
E3	2.5 /6 (41.6%)	2.8 /6 (46.7%)	73	**-5.1**	31.4	26.3
E4	3.0/6 (50.0%)	3.1 /6 (51.6%)	74	**-1.6**	24	22.4
E5	4.0 /6 (66.7%)	3.5 /6 (58.3%)	76	8.4	9.3	17.7

E6	2.1 /6 (35.0%)	2.6 /6 (43.3%)	72	**-8.3**	37	28.7
E7	3.5 /6 (58.3%)	4.8 /6 (80.0%)	78	**-21.7**	19.7	**-2**
E8	4.2 /6 (70.0%)	3.9 /6 (65.0%)	82	5	12	17
E9	3.3 /6 (55.0%)	4.6 /6 (76.7%)	77	**-21.7**	22	0.3
E10	2.5 /6 (41.6%)	2.3 /6 (38.3%)	77	3.3	35.4	**-38.7**

In this study, GL scores were used as the accepted scores because they are based on the average scores of 140 language lecturers. Therefore, both scores from individual lecturers (IL) and 'PaperRater.com' (PR) were compared to the scores produced by the group lecturers (GL) to determine the difference. Overall, the findings show a range of differences among the total scores marked by IL, GL and PR. The smallest difference is between IL and GL which is -1.6 and biggest difference is between GL and PR which is 38.7.

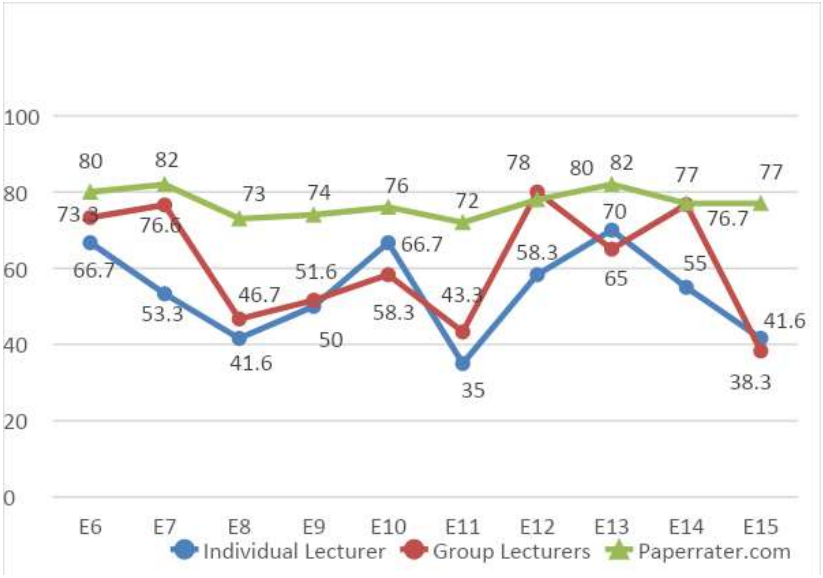

Figure 1: Trends of Scores of Essays

Based on Figure 1, out of 10 essays, only one score from IL (E4) matches with GL score. Similarly, only one score by PR that matches GL score (E9). Generally, it can be concluded that PR scores, on average (mean = 85.1), are more lenient compared to papers marked by IL and GL. This is based on the average scores of three types of raters to which PR recorded an average score of 85.1, GL=61.82 and IL=77.1.

4.2 The Correlation between the Scores of Essays Marked by Individual Lecturers and Group of Lecturers and 'PaperRater.com'

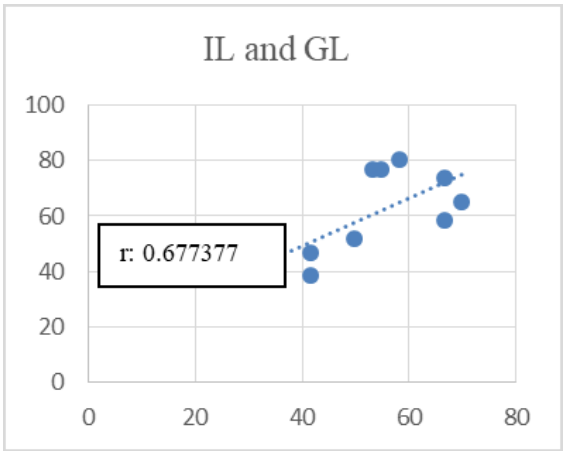

Figure 2: Correlation between IL and GL essay scores

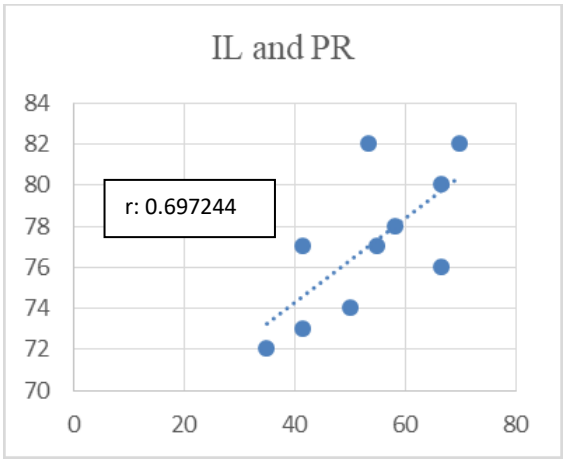

Figure 3: Correlation between IL and PR essay scores

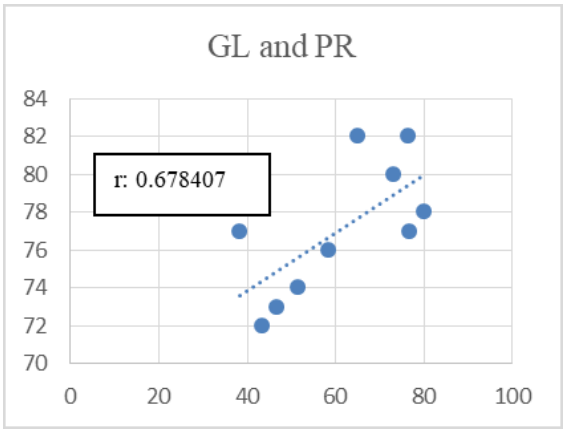

Figure 4: Correlation between GL and PR essay scores

To further investigate the relationship of scores among the three types of raters, a correlation analysis was conducted. The correlation is found to be moderate (refer to Table 3). Thus, it can be concluded that there is a moderate positive correlation among the three IL, GL and PR (r values = 0.677377, 0.697244 and 0.678407 respectively). The difference among the positive correlation values of the three types of markers is small.

4.3 The Extent to Which Paperrater Can Be Used as a Viable Assessment Tool

Table 8: Results and Status of all Essays

Essay	IL Score	Grade	GL Score	Grade	PR Score	Grade	Result
E1	66.7	B	73.3	B+	80	A	Rejected
E2	53.3	C	76.6	A-	82	A	Accepted
E3	41.6	D	46.7	C-	73	B+	Rejected
E4	50.0	C	51.6	C	74	B+	Rejected
E5	66.7	B	58.3	C+	76	A-	Rejected
E6	35.0	E	43.3	D	72	B+	Rejected
E7	58.3	C+	80.0	A	78	A-	Rejected
E8	70.0	B+	65.0	B	82	A	Rejected
E9	55.0	C+	76.7	A-	77	A-	Accepted
E10	41.6	D	38.3	E	77	A-	Rejected

Even though there is a moderate positive correlation among IL, GL and PR, it is essential to determine whether PR is a viable assessment tool for marking essays. Therefore, the total scores and grades were compared in order to determine whether they fall in the accepted range (Table 8). Scores were determined to be accepted, as described earlier, when the difference among the raters did not affect the grade stipulated by the university's grading guide (Table 6).

Table 8 shows the summary of the comparison made. It was found that only 2 out of the 10 essays (E2 and E9) in this category were found to fall within tolerable range of acceptance where it recorded a minimal value of 0.3% difference (GL 76.7% and PR 77%) and both assessments remained to be in the same grade. The rest of the essays recorded considerable differences ranging from 2% to 28.7%.

5. Discussions

The overall findings indicate that PR is more lenient than human raters. Even though there exist unacceptable gaps among the scores from both sides in Table 7, it is important to notice that the score patterns of E1 to E10 for GL and PR are quite similar. This is proven by the correlation value which recorded a moderate positive relationship ($r = 0.67$). This means that PR is able to identify the quality of those essays but the only difference is that PR is seen to be more lenient with the marks. The leniency of PR is seen through the mean scores of the ten essays of which PR recorded 81.5 marks while GL's mean score is 61.82. This is similar to the finding by Wang and Brown (2007) where they concluded that the mean score given by the AES is higher than human raters'.

The findings also show that there exists inconsistency in human essay rating (IL and GL). When compared IL and GL, the range of differences is greater and the correlation is the lowest ($r=0.677$). It is obvious that IL is stricter since there are 7 out of 10 essays were marked lower than GL (Table 8). Severity and leniency in marking are

commonly attributed as human-rater errors and biases (Mo Zhang, 2013). Thus, the scores by PR are more in line with GL, however, more lenient than IL as illustrated in Figure 1.

Next, the findings indicate that there is a tendency for PR (AES) to be more lenient compared to IL and GL (Human raters). Referring to Table 7, PR scored only one essay (E7 essay) lower than GL with a difference of 2 marks. The other scores were higher than GL ranging from 0.3 to 38.7 marks. An extreme difference between GL and PR is indicated in essay E10. GL marked E10 as 38.3 marks while PR marked as 77. This maybe contributed by the inability of PR to assess relevance, context and meaning of the writing task. The PR version used in the study is basically restricted to assess only sentence structure and readability (Table 4).

6. Conclusions

With regards to the viability of as an assessment tool, PR can be a useful tool for language lecturers. Although it has been proven through this study that there are major differences in terms of overall assessment between lecturers and PR, the usability of such system should not be totally ignored. Its usefulness in assisting language instructors in managing students' essays is still relevant and beneficial. It can help to reduce the bottleneck issue faced by lecturers which happens when students submit their essays at the same time and it can be overwhelming to lecturers. Assessment from 'PaperRater.com' can become the first layer of assessment before the writing is finally assessed by the lecturer. In this way, the students will be able to independently check for more minute aspects of assessment like vocabulary check, plagiarism, grammar and spelling check. These are aspects that can help language instructors to better handle the task of examining students' essays.

Apart from this, 'PaperRater.com' through automated scoring system will enable students to get instant feedback, a quality that is closely related to increasing motivation among students as proven in many other previous studies (Somaye Toranj and Dariush Nejad Ansari, 2012; Ng Sing Yii et al., 2016). Besides, each section of 'PaperRater.com' report is presented in both score figures and quite detail explanation of each section, for instance, in reporting vocabulary words, 'PaperRater.com' presented the report by stating the score and suggest the level to which the writer need to achieve in order to improve. Results of the present study have brought about implications for future research. Research on students' and lecturers' feedback on AES as an assessment tool is recommended to find out its benefits to the users. Investigations on whether AES affects students' motivation and learner autonomy are also worth considering. This will help learners to enhance their self-regulation skills in performing writing tasks. Besides, the use of 'PaperRater.com' in assessing essays in classroom will give an alternative approach of assessment for the lecturers and more importantly, the students will be able to benefit more from this exercise.

Finally, findings of the study can assist AES developers to improve their tools since lecturers and students are their end-users. The main objective may not only linger on the reliability and consistency of scoring but more challenging issues in the absence of human rater in considering elements that involve inferential skills, critical thinking and abstract ideas presented by the students in their essays. A new approach or combination of approaches used in the AES system may need to be reconsidered apart from what have been commonly used such as the Natural Language Processing (NLP) technique, statistical approach, discourse structure analysis, syntactic structure analysis, vocabulary usage analysis and also the corpus-based approach in order to further improve AES in the future.

Apart from this, future research should also consider the scale of the study where more essays are needed to be selected as samples. Small number like 10 essays selected in this study may not be able to capture the various style and level of the writer thus making any generalization about the findings to be very much restricted. This, however, will require substantial amount of fund and time too.

Acknowledgements

The researchers would like to thank the Akademi Pengajian Bahasa, UiTM for supporting this research.

About the Authors

Mohammad Radzi Manap (PhD) is a senior lecturer at the department of English and Linguistics at the Akademi Pengajian Bahasa, Universiti Teknologi MARA with more than 20 years of experience as instructors of various English language and linguistics courses at tertiary level. His research interests include the use of technology in language teaching, TESL and sociolinguistics.

Nor Fazlin Mohd Ramli (PhD) is a senior lecturer at the department of English and Linguistics at the Akademi Pengajian Bahasa, Universiti Teknologi MARA who has great interest in the field of technology in language learning, instructional design as well as TESL.

Aini Akmar Mohd Kassim (PhD) is a senior lecturer at the department of English and Linguistics at the Akademi Pengajian Bahasa, Universiti Teknologi MARA. Her areas of interest include L2 teaching and learning, learning strategies and learner strategies as well as learner autonomy.

References

Attali, Y. & Burstein, J. (2006). Automated Essay Scoring with e-rater® V.2. *Journal of Technology, Learning and Assessment, 4(3)*. Available from http://www.jtla.org

Creswell, J. (2009). Research design: Qualitative, quantitative, and mixed methods approaches (3rd ed.). Thousand Oaks, CA: Sage.

Dikli, S. (2006). Automated Essay Scoring. *Turkish Online of Distance Education*. Vol. 7 (1). pp 49-62.

Grimes, D. & Warschauer, M. (2010). Utility in a Fallible Tool: A Multi-Site Case Study of Automated Writing Evaluation. *Journal of Technology, Learning, and Assessment*,*8*(6). Retrieved [22.5.2016] from http://www.jtla.org.

Hinkle D.E, Wiersma W, Jurs. (2003). 'Applied Statistics for the Behavioral Sciences'.

Kheradmand, N. and Sayadiyan, S. (2016). Comparative Investigation of the Effects of Immediate and Delayed Error Correction on the Achievement of Male and Female Iranian EFL Learners' Writing Skill. *International Journal of Social Science and Education*. Vol. 6 (1). Retrieved from www.ijsse.com/sites/default/files/issues/2016/v6i1/paper-04.pdf [21 July 2015]

Moore, D. S., Notz, W. I., & Flinger, M. A. (2013). The basic practice of statistics (6th ed.). New York, NY: W. H. Freeman and Company.

Mo Zhang. (2013). Contrasting Automated and Human Scoring of Essays. R&D Connections, 21. Retrieved [11.6.16] from 2013 https://www.ets.org/Media/Research/pdf/RD_Connections_21.pdf

Nafizah Hamidun, Shafiq Hizwari Md Hashim and Nur Farhinaa Othman. (2012). Enhancing Students' Motivation by Providing Feedback on Writing: The Case of International Students from Thailand. *International Journal of Social Science and Humanity*. Vol. 2 (6). Retrieved [21 July 2016] from www.ijssh.org/papers/179-A10062.pdf

Neda Kheradmand and Sima Sayadiyan. (2016). Comparative Investigation of the Effects and Delayed Error Correction on the Achievement of Male and Female Iranian EFL Learners' Writing Skill. *International Journal Soc. Sci. & Education*. Vol. 6 (1). Retrieved [21 July 2016] from www.ijsse.com/sites/default/files/issues/2016/v6i1/paper-04.pdf

Ng Sing Yii, Bong Chih How, Lee Nung Kion and Hong Kian Sam. (2016). Automated Essay Scoring Feedback (AESF): An Innovative Writing Solution to the Malaysian University English Test (MUET). *International Journal on E-Learning and Higher Education*, 4. pp. 130-143.

Qiu Yubing. (2016). Pigai Smart Essay Scoring System and Its Implications for Teaching English Writing. *Journal of Applied Science and Engineering Innovation*, Vol. 3 (6). pp. 217-219. Retrieved [2 February 2018] from http://www.jasei.org/PDF/3-6/3-217-219.pdf

Saadiyah Darus, Siti Hanim Stapa, Supyan Hussin & Koo Yew Lie. (2000). A Survey of Computer–based Essay Marking (CBEM) Systems. *International Conference "Education & ICT in the New Millenium'*. pp 519-528. Retrieved [14.3.2019] from https://www.academia.edu/6977567/A_survey_of_computer_based_essay_marking_CBEM_system

Saadiyah Darus, Siti Hamin Stapa & Supyan Hussin. (2003). Experimenting A Computer-Based Essay Marking System At Universiti Kebangsaan Malaysia. *Jurnal Teknologi*, 39(E) Dis. 2003: 1–18. Retrieved [23.5.2016] from

http://www.academia.edu/263332/Experimenting_a_Computer-Based_Essay_Marking_System_at_Universiti_Kebangsaan_Malaysia

Shermis, M.D., Burstein, J., Higgins, D., and Zechner, K. (2010). Automated essay scoring: Writing assessment and instruction. In E. Baker, B. McGaw, and N.S. Petersen (Eds.), *International encyclopedia of education* (3rd ed., pp 75-80). Oxford, England: Elsevier.

Somaye Toranj and Dariush Nejad Ansari. (2012). Automated Versus Human Essay Scoring: A Comparative Study. *Theory and Practice in Language Studies*, Vol. 2, No. 4, pp. 719-725, April 2012

Uzun, K. (2018). Home-Grown Automated Essay Scoring in the Literature Classroom: A Solution for Managing the Crowd? *Contemporary Educational Technology,* 9(4). pp.423-436. Retrieved [15.3.19] from http://doi.org/10.30935/cet.471024

Valenti, S., Neri, F., and Cucchiarelli, A. (2003). An Overview of Current Research on Automated Essay Grading. *Journal of Information Technology Education. 2*. pp 319-330.

Wang, J. & Brown, M. S. (2007). Automated Essay Scoring Versus Human Scoring: A Comparative Study. *Journal of Technology, Learning, and Assessment*, 6(2). Retrieved [25.5.2016] from http://www.jtla.org

Warschauer, M., & Grimes, D. (2008). Automated writing assessment in the classroom. Pedagogies: An International Journal, 3, pp 22–36. Retrieved [3 February 2018] from http://education.uci.edu/uploads/7/2/7/6/72769947/awe-pedagogies.pdf.