Journal of
Creative Practices in Language Learning and Teaching (CPLT)
Volume 5, Number 1, 2017

Automated Essay Assessment: An Evaluation on *PaperRater*'s Reliability from Practice

Nguyen Vi Thong
vithong1985@gmail.com
Graduate Institute of Linguistics
National Chung Cheng University, TAIWAN

Received: 3 Mar 2017. Accepted: 22 Apr 2017 / Published online: 11 May 2017

ABSTRACT
From a perspective of a PaperRater user, the author attempts to investigate the reliability of the program. Twenty-four freshman students and one writing teacher at Dalat University - Vietnam were recruited to serve the study. The author also served as one scorer. The scores generated by PaperRater and the two human scorers were analyzed quantitatively and qualitatively. The statistical results indicate that there is an excellent correlation between the means of scores generated by three scorers. With the aid of SPSS and certain calculation, it is shown that PaterRater has an acceptable reliability which implies that the program can somehow assist in grading students' papers. The semi-structured interview at the qualitative stage with the teacher scorer helped point out several challenges that writing teachers might encounter when assessing students' prompts. From her perspective, it was admitted that with the assistance of PaperRater, the burden of assessing a bunch of prompts at a short time period would be much released. However, how the program can be employed by teachers should be carefully investigated. Therefore, this study provides writing teachers with pedagogical implications on how PaperRater should be used in writing classrooms. The study is expected to shed new light on the possibility of adopting an automated evaluation instrument as a scoring assistant in large writing classrooms.

Keywords: *PaperRater*. Writing assessment. Automated essay evaluation. Reliability. Human scorers

✉ Nguyen Vi Thong
Graduate Institute of Linguistics
National Chung Cheng University, TAIWAN
E-mail: vithong1985@gmail.com

INTRODUCTION

In English as a Second or Foreign Language (ESL/EFL) writing learning, it has been widely agreed that more practice can benefit students' writing skills (e.g., National Commission on Writing in America's Schools and Colleges, 2003, p. 3). This requirement may lead writing teachers to an overburdening situation in that they do not have sufficient time to mark students' papers (Warschauer and Grimes, 2008). The authors also claim that with the assistance of Automated Writing Evaluation (AWE; also called automated essay assessment or scoring), which uses artificial intelligence to score and assess essays, teachers can be set free from this burden, thus it can encourage more writing practice and faster improvement. This paper aims to investigate the newly-developed free online AWE called *PaperRater* as an instance to examine whether AWE really works in writing classroom.

AWE TOOLS AND SURROUNDING ARGUMENTS

The idea of developing automated writing evaluation programs began in the years of 1960s when a scoring model based on a corpus of essays previously graded by hand was built to measure the essay length and average sentence length (Shermis, Mzumara, Olson, & Harrington, 2001). During 1990s, two considerable competing automated essay scoring engines named E-rater® and Intellimetric were developed by Educational Testing Service and Vantage Learning, which opened a new age of web-based language testing (Burstein, 2003; Elliot & Mikulas, 2004). In Attali and Burstein (2005), it is reported that these programs are developed to assess learners' writing skills and provide them with instantaneous score reporting and diagnostic feedback. More than 50 features are created as criteria to predict the essay's core such as grammar, vocabulary, style, organization, development, lexical complexity, essay length, etc. Simultaneously, a group of academics developed another writing assessment tool named Intelligent Essay Assessor, which used latent semantic analysis to score essays. This technology allowed the semantic meaning of the piece of writing to be compared with a boarder corpus of textual information on a similar topic (Landauer, Laham, & Foltz, 2003).

Whenever an automated scoring tool is utilized, its reliability should be taken into careful consideration (Warschauer and Grimes, 2008). In reality, the reliability of the aforementioned AWE tools has been extensively examined by comparing the correlations between computer-generated and human-rater scores with the correlations attained from two human raters (Cohen, Ben-Simon, & Hovav, 2003; Keith, 2003). The results show that AWE score agrees roughly with a human-rated score more than 95% of the time, which is also the same rate of correlations between two human scorers. This figure led the organization to a decision to bring these tools to commercial markets when they are employed to score writing papers in TOEFL iBT.

Eventually, those three AWE engines mainly serve commercial purposes; benefits toward classroom teachers and learners are not very reachable. Moreover, in Chapelle and Douglas (2006), it is emphasized that the combination of language assessment and technology provides the potential to efficiently strengthen computer-based tasks. Therefore, recently more attention has been paid to development of AWE tools to be directly used in classroom. As listed in Warschauer and Grimes (2008), there have been several pieces of software on AWE for

classroom such as Criterion developed by ETS Technology, My Access by Vantage Learning, and WriteToLearn by Pearson knowledge Technologies. Each program combines the functions of scoring with the provision of grammar, spelling, mechanical feedback, and a range of support resources. However, it seems to be problematic when the software can score students' writing prompts only if the prompts come with the program. It means that all the students must have one of these programs in hand, which seems not to be perfect to all students.

In fact, Shin (2012) stresses that even though web-based language testing may "enhance test authenticity and reliability by making possible a rich contextualized input, various response formats, and automated scoring", there has still been very little study conducted to investigate whether online testing can actually works in writing classroom (p.277). This situation motivates the writer of this project to investigate *PaperRater*.com, which is a free resource utilizing Artificial Intelligence to help learners write better and teachers score papers faster. *PaperRater* is a combination of Natural Language Processing, Machine Learning, Information Retrieval, Computational Linguistics, and Data Mining to create a powerful automated proofreading engine available online. Especially, *PaperRater* does not require students to use prompts run by the program; students can upload their Microsoft Word file to the page. After analyzing the genre-categorized text, the program will provide the user with a clear range of assessment features, feedbacks and scores. The features that the program sets as criteria to assess texts differ for selected genres, levels of learners. However, in general, a text is evaluated from various features such as grammar, spelling, (academic) vocabulary variety, transitional words, style, plagiarism detection, etc. Finally, the program offers an overall score for the whole text totally integrated from the score of each feature above.

Although as promoted on its page (www.paperrater.com) that *PaperRater* has been widely used in numerous countries, to my knowledge, there has not been any study intensively investigating this online resource.

Research questions

As what described above, *PaperRater* is worth an investigation to determine whether it should be employed in classroom. In the scale of this project, the writer only pays attention to how *PaperRater* helps language teachers in terms of assessing and scoring writing assignments. Therefore, the paper attempts to answer the following questions:

1. What is the reliability of *PaperRater*? What is the correlation between *PaperRater*-generated and human-rater scores?
2. Where do the differences lie between *PaperRater*-based assessment and human-based assessment?
3. What pedagogical implications can be derived from this investigation?

METHOD

Participants

The participants in this study were 24 Vietnamese undergraduate students majoring in English.

They were currently taking the *Writing 1* course at Dalat University (DLU) in Vietnam. All of them were freshmen in their program, and they voluntarily participated in the study. It is important to consider the English levels of participants since *PaperRater* scores the papers based on the levels set by the program. Therefore, "*undergraduate students*" should be selected as "*education level of the author*" before the assessment is run.

Moreover, the teacher who was instructing the course *Writing 1* at the time at DLU was willing to be the first scorer for the papers. The second scorer was the author of this study. The papers were evaluated based on the criteria set by *PaperRater* in the first stage, and then they would be scored based on the rubric developed by the author.

Procedures

In order to address the first research question, the participants were required to write a 200-word essay, topic of which was related to a favorite holiday or festival. They were encouraged to compose their prompts with *Text document* so as to prevent them from using the function of *Spelling & Grammar checking* in *Microsoft Word*. It should be noted that *PaperRater* evaluates the papers based on their genres. Therefore, for this set of data, "*essay*" should be selected as the target genre. The scores for prompts generated by *PaperRater* were then recorded. It should also be noted here that there is no clear rubric generated by *PaperRater;* this program only suggests several categories which are used to assess the prompts, and based on these categories, Rubric 1 (Appendix A) was developed as an instrument for teacher scorers to assess the students' papers. Simultaneously, the prompts were scored by the teacher scorers depending on the criteria set in Rubric 1 in the first stage and Rubric 2 (Appendix 2) in the second stage. The analysis for the scores generated by *PaperRater* and the teacher scorers would be processed through two stages: quantitative and qualitative analysis.

Quantitative analysis

In this stage of analysis, the teacher scorers relied on the criteria set in Rubric 1 and grade the students' prompts. This rubric was basically designed based on the main features that *PaperRater* utilizes to assess the papers. The scores were then compared to those generated by *PaperRater* with the assistance of SPSS to define the inter-rater reliability and the correlation among variables. In concrete, the comparison was going through the following steps:

1. *Internal-reliability*: *Intraclass Correlation Coefficient (ICC)* was run to assess the inter-rater reliability of the grades. The result of this measurement can imply the consistency and agreement of the data-set.
2. *Descriptive statistic*: The purpose of this step was to figure out several basic features of the variables such as mean, median, range, standard deviation, maximum, minimum, etc.
3. *ANOVA*: With the null hypothesis (H_0) that the means of these three sets of score are equal and H_1 assuming that at least one of the means is different, ANOVA is performed to check the differences between the means of the three populations. If

there is any difference, it means that the null hypothesis is rejected. The results of this test can imply the reliability of *PaperRater*'s assessment.

Sample paper assessment analysis and Interview

In order to serve this manual analysis, top 5 prompts with the highest deviation between the scores generated by *PaperRater* and teacher scorers were chosen as samples. The selection of papers to be the samples was based on the result from the formula:

$$D = S_p - (S_1+S_2)/2$$

in which

- D: Deviation
- S_p: Score assigned by *PaperRater*
- S_1: Score assigned by scorer 1
- S_2: Score assigned by scorer 2

It implies from the formula that the scores assigned by *PaperRater* were compared to the average of scores generated by the two teacher scorers. By this calculation, top 5 papers with the highest deviation were decided as samples to be further examined. This analysis could help in finding the differences between computerized and human evaluations. This also might suggest advantages and disadvantages of these two means of scoring. The finding could imply which points each scoring method might be missing so that a further suggestion could be made on how much we can rely on each.

In the next stage, it can be clearly seen that there are some aspects of writing assessment that computerized tools may overlook such as evaluating the content, organization, and main idea of the writing prompt; therefore, the author developed Rubric 2 to assess thoroughly all the features that writing evaluation should go through. The results of this step could imply how the scores changed after Rubric 2 was applied. It is expected that a conclusion of how *PaperRater* can be used by writing teachers in terms of assessing and scoring should be drawn from the above findings. It means that an answer of how *PaperRater* can be combined in classroom should be thoroughly jotted down.

However, in order to avoid the subjectiveness for any assumption relating to the results, it was decided to conduct a semi-structured interview with scorer 1. The results from the interview was expected to provide an evaluation having been accumulated from the real classroom context.

RESULTS

This section aims to find the answers for three main questions: the reliability of *PaperRater*, the difference between two scoring methods, and the pedagogical implications. Simultaneously, discussions and assumptions upon the findings will be also presented.

Statistic-based evaluation

The first statistical analysis was conducted to estimate the *Intraclass Correlation Coefficient* to

measure the internal consistency of the data-set. Three items of scores generated by *PaperRater* and the scorers were manipulated. As can be seen in Table 1, the average measures equal 0.742, which assumes that the level of internal reliability of the data-set is approximately *excellent* level (.75). It means that within the data-set, there is an absolute correlation between the items.

Table 1
Intraclass Correlation Coefficient

		95% Confidence Interval	
	Intraclass Correlation[b]	Lower Bound	Upper Bound
Average Measures	.742	.492	.880

The above results can be a good preparation to continue on the analysis to the step of assessing the reliability of *PaperRater*. It is necessary here to recall the definition of reliability in language assessment. As discussed in Hossein (2012), reliability can be technically understood as "the extent to which a test produces consistent scores at different administrations to the same or similar group of examinees" (p.39). Hence, to evaluate the reliability of *PaperRater*, two correlations were examined: the first one is between scores assigned by the two teacher scorers and the second one is between scores generated by *PaperRater* and the two teacher scorers based on the Rubric 1. Figure 1 and Table 2 provide a statistical overview and description of the scores in terms of quantitative analysis.

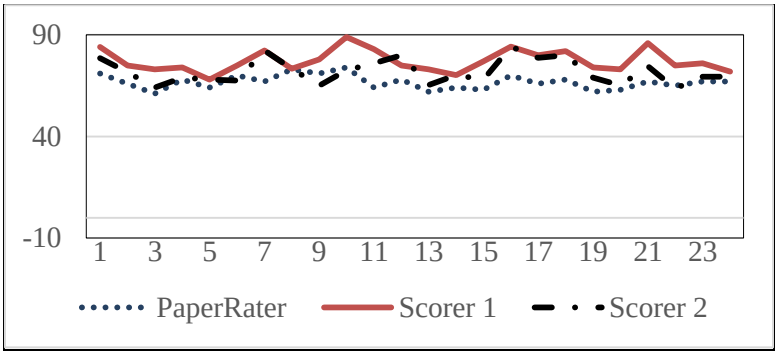


Figure 1. Scores after Rubric 1

Table 2
Descriptive Statistic

	PaperRater	*Scorer 1*	*Scorer 2*
Mean	66.71	77.18	71.92
Median	67.00	75.00	69.80
Mode	67.00	75.00	69.00
Standard Deviation	3.53	5.41	6.11
Minimum	61.00	68.00	63.60
Maximum	74.00	89.00	84.20

Count	24.00	24.00	24.00

Figure 1 shows the whole picture of how the prompts are scored by the software and human scorers. It can be clearly seen that there is quite a great difference in the assigned scores, not only between computer and human scorers but even between the two human scorers. Table 1 then provides a clear comparison of the assigned scores. From the information on the figure and table, it can be assumed that all of the scorers generate the scores in a relatively consistent way. In most cases, *PaperRater* assigns the lowest scores, whereas scorer 1 constantly generate the highest scores and scorer 2 stands in the second position. This trend is reflected by their score means of 66.71; 77.18; and 71.92, respectively. This situation is also shown through their medians, modes, and sums.

However, in order to check whether there is a significant difference between the means of three grading methods, ANOVA was performed in this stage. The result of this step is shown is Table 3.**:**

Table 3
ANOVA

Source of Variation	*SS*	*df*	*F*	*F crit*
Between Groups	1314.623	2	24.92	3.12
Within Groups	1819.777	69		
Total	3134.4	71		

As can be seen in Table 3, the *F* value is much greater than *F critical* value, so it can be assumed that the null hypothesis is rejected. It means that the means of the three populations are not all equal. This result implies that although there is an excellent correlation between the variables, there is still a disparity between their means. The matter concerned at this point is how to examine the real reliability of *PaperRater.*

Therefore, it comes up to a decision of how to measure the reliability of *PaperRater.* This decision derives from the observation that there is a disparity not only between the human scorers and *PaperRater* but also within the two human scorers, and that the scores given by scorer 2 mostly stand at the point between those given by *PaperRater* and scorer 1. Therefore, the mean score generated by scorer 2 was chosen to be a reference point, from which its disparity with the mean scores assigned by *PaperRater* and scorer 1 were compared. The results of this measurement show that the deviation between the scorer 2 and *PaperRater* equals ±5.21; and ±5.26 for scorer 2 and scorer 1. It can be seen that the disparity seems to be roughly identical. This calculation can imply that in terms of consistency and disparity, and at this point *PaperRater* proves its reliability.

Hence, in the aspect of quantitative analysis, *PaperRater* can be reliable enough to be used as a writing assessment instrument. However, where the disparity lies between the instrument and human scorers will be investigated in the next stage.

Sample paper analysis

By applying the formula presented in the section of method, 5 sample papers were sorted out for further investigation. These sample papers were written by students numbered 1, 11, 12, 17, and 21, deviations of whose scores can be easily seen in the Figure 1.

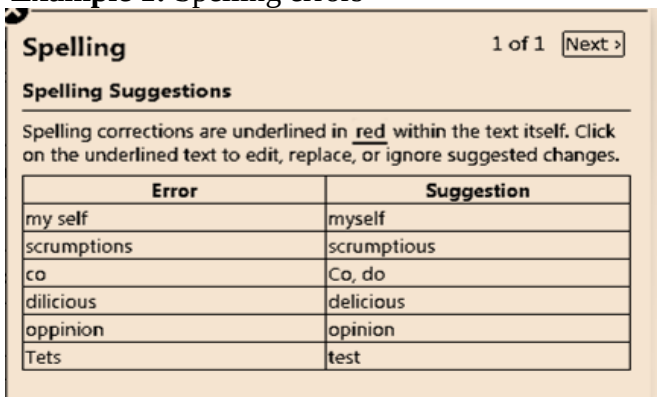
As can be seen in Rubric 1, there consist of 5 categories that both *PaperRater* and teacher scorers used to assess the prompts. For the first two categories (*spelling* and *grammar*), a comparison in the number of errors detected by each scorer was conducted. In contrast, for the category of *vocabulary words*, the number of academic words were counted; it means that the more academic or difficult words are found in the prompt, the higher the score can be given. The results are shown in Table 4.

Table 4
Assessment of three categories

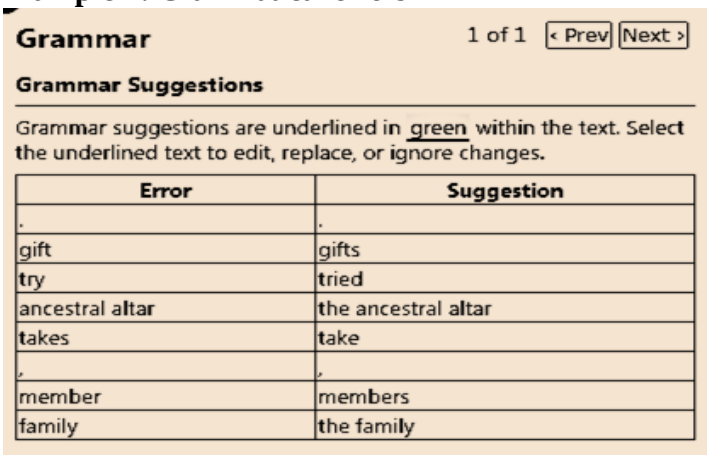
	Spelling (errors)			Grammar (errors)			Vocabulary words (amount)		
Student	S_p	S_1	S_2	S_p	S_1	S_2	S_p	S_1	S_2
1	8	3	5	4	3	4	6	5	6
11	6	2	4	3	2	3	3	4	3
12	5	4	2	3	4	3	2	5	4
17	9	4	5	4	1	2	3	4	3
21	6	3	8	4	2	4	5	5	3
Total	**34**	**16**	**24**	**18**	**12**	**16**	**19**	**23**	**19**

Table 4 reveals the ability to detect errors and academic words of *PaperRater* and the two teacher scorers. This result also partly explains why the papers were scored differently. It can be seen that in terms of detecting spelling and grammar errors, *PaperRater* performs better, whereas the scorer 1seems to detect the fewest errors. As observed, most of the errors found by *PaperRater* tend to be related to tenses, articles, number features, punctuation, etc. In other words, the program can effectively figure out basic grammatical and spelling errors. The following examples show how *PaperRater* detects the errors and suggests the corrections.

Example 1: Spelling errors



Example 2: Grammatical errors



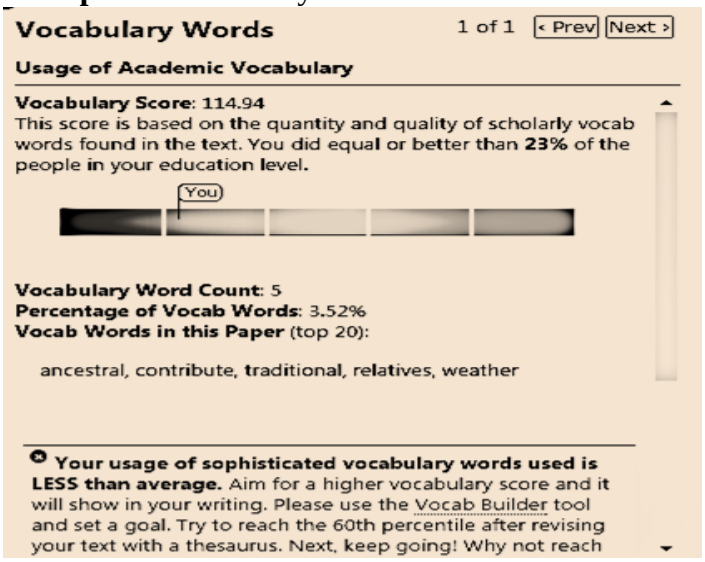
Grammar 1 of 1 ‹ Prev Next ›

Grammar Suggestions

Grammar suggestions are underlined in green within the text. Select the underlined text to edit, replace, or ignore changes.

Error	Suggestion
.	.
gift	gifts
try	tried
ancestral altar	the ancestral altar
takes	take
,	,
member	members
family	the family

Example 3: Vocabulary words



Vocabulary Words 1 of 1 ‹ Prev Next ›

Usage of Academic Vocabulary

Vocabulary Score: 114.94
This score is based on the quantity and quality of scholarly vocab words found in the text. You did equal or better than **23%** of the people in your education level.

You

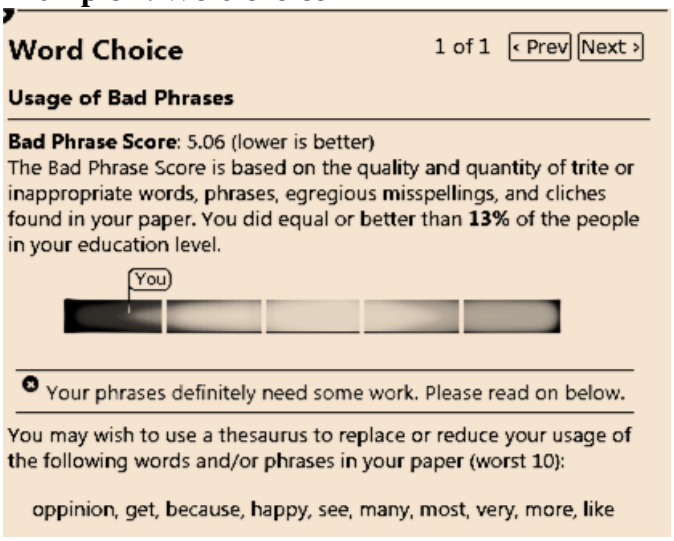
Vocabulary Word Count: 5
Percentage of Vocab Words: 3.52%
Vocab Words in this Paper (top 20):

ancestral, contribute, traditional, relatives, weather

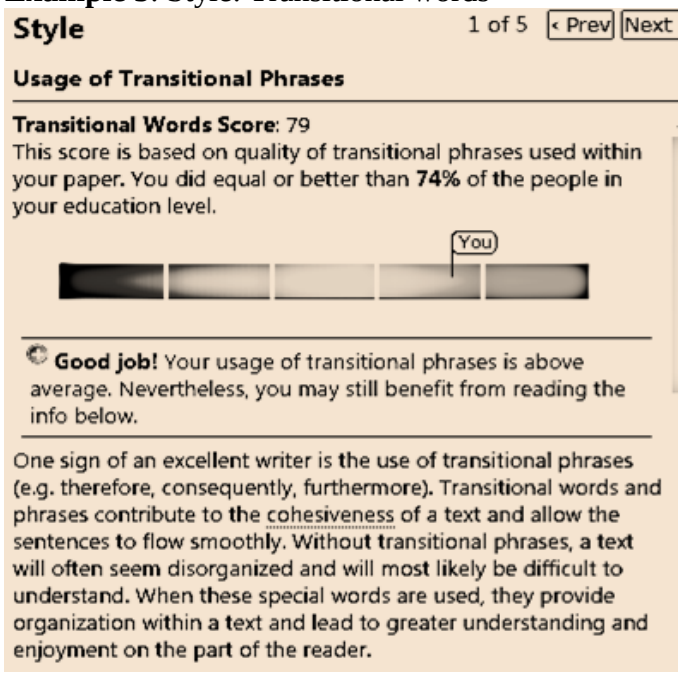
Your usage of sophisticated vocabulary words used is LESS than average. Aim for a higher vocabulary score and it will show in your writing. Please use the Vocab Builder tool and set a goal. Try to reach the 60th percentile after revising your text with a thesaurus. Next, keep going! Why not reach

Regarding the two rest categories in Rubric 1 (*word choice* and *style*), it became rather difficult for the author to evaluate how the human scorers graded these categories in the prompts since the scorers only suggested the final scores for each criterion without leaving any notes or comments. However, *PaperRater* has been programmed to detect and count some criteria relating to these two categories such as *number of bad phrases*, *sentence length*, *sentence beginning*, *transitional words*, and *passive voice*. The examples below show how these categories are assessed by *PaperRater*.

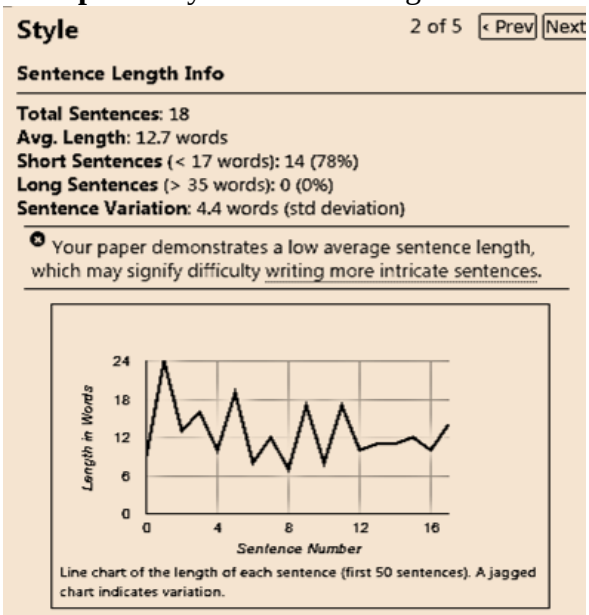
Example 4: Word choice



Example 5: Style: Transitional words



Example 6: Style: Sentence length



As can be seen in the examples above, *PaperRater* can carefully assess students' prompts with the exact statistics, whereas regarding these categories, teacher scorers seem basically to depend on observation and intuition. It could be really time-consuming to do the same statistics as what the program does. Therefore, as what has been analyzed so far, it can be assumed that in terms of grading the categories in Rubric 1, *PaperRater* can give a better performance. This can be a basis for classroom implications which will be further discussed in the next section.

Rubric 2 and Pedagogical implications

As widely known by writing teachers, Rubric 1 is not effective enough to evaluate a writing prompt; it needs more categories which *PaperRater* cannot cover yet. For this reason, Rubric 2 was designed to evaluate the prompts thoroughly. The following figures illustrate how the scores assigned by the teacher scorers changed after applying the Rubric 2.

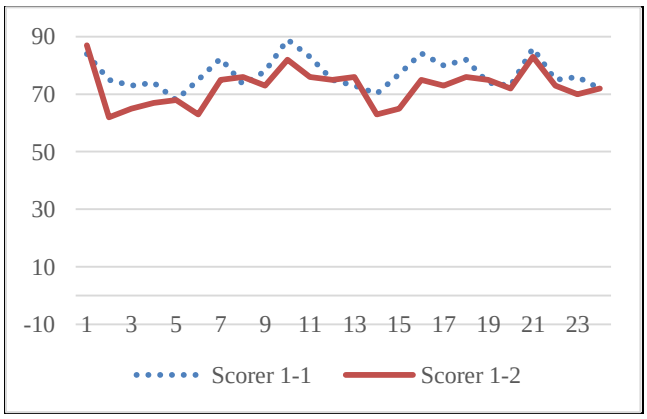


Figure 2. Scorer 1: Scores after Rubric 1 and Rubric 2

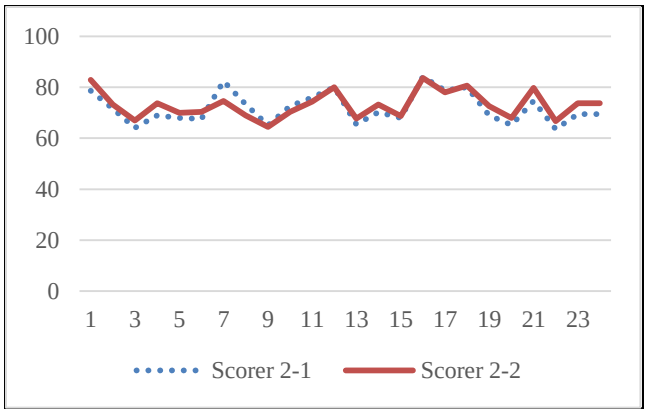


Figure 3. Scorer 2: Scores after Rubric 1 and Rubric 2

Table 5
Descriptive Statistics after Rubric 2

	Scorer 1		*Scorer 2*	
	Rubric 1	*Rubric 2*	*Rubric 1*	*Rubric 2*
Mean	77.2	72.6	71.9	73.2
Standard Deviation	5.4	6.4	6.1	5.3

It can be seen in the figures above that the scores generated by the two human scorers, especially in the case of scorer 2, considerably change after Rubric 2 is employed. In concrete, Table 5 provides the difference in means and standard deviations of the scores generated by each teacher score before and after Rubric 2 was employed. It is relatively clear that the difference is noticeable, especially for the case of score 1.The figures and table above also imply that even though *PaperRater* can quite effectively assess the prompts based on the Rubric 1, it cannot totally replace human scorers since a lot of important criteria in writing assessment may be ignored. Therefore, it raises the question why and how *PaperRater* can be used in writing classrooms.

As analyzed so far, even scores between two human scorers can be as different as those between one and *PaperRater*. To find the explanation for this situation, a semi-structured interview with scorer 1 was conducted. The interview included three prepared-in-advance questions relating to the rubric design, writing scoring method, and automated essay assessment. During the interview, some more open questions were raised according to the answers of the interviewee, which was expected to provide a deeper view toward the issue. The result of the interview can be jotted down within the following confirmations:

- The rubrics are effective, but it is still difficult for her to identify the terms "*few*", "*many*", "*sometimes*", "*rarely*", or "*hardly*". It means that in most of the time, she used her own intuition and judgement. This situation raises a requirement to design an ideal rubric for assessing writing, in which the concrete statistics should be

mentioned. However, in reality this can become impossible to writing teachers since it would take really much time.

- In her opinion, intuition plays an important role in assessing writing. She normally does not have any rubric when grading students' homework. In this case, intuition can lead her to the decision of which grades each prompt can receive. This phenomenon can explain why the personality of writing teachers can influence the scores for writing works.
- Correcting every error in students' prompts can be burdening to her. In reality, she has to instruct several writing classes at the same time, each of which comprises at least 40 students. With less important pieces of writing homework, she does not often correct all the students' errors. Therefore, she agrees that if there is any assistance of computerized assessment, it would be very convenient and time-saving.

From the interview, it can be implied that even though there is a reference from the rubrics, it can be still difficult for teachers to assess writing prompts. Therefore, it would be ideal for the teachers if there is a tool which can partly help them in dealing with scoring the students' assignments. From what has been analyzed above, *PaperRater* should be highly recommended to be an assistant in writing classrooms. Certainly, it can be used totally in scoring papers, but it should be combined in an appropriate way.

DISCUSSION AND CONCLUSION

The results of the study is one more approval for the claim of Chapelle and Douglas (2006) that computerized teaching tools and technologies should be effective aids in language classrooms. It can be seen from the result of the statistical analysis that the scores generated by *PaperRater* and the two human scorers are considerably consistent. It means that each subject of scorer assesses the set of students' prompts in a stable manner. Moreover, it is interestingly found that the difference in scores takes place not only between the program and each human scorer but within the two human scorers. This situation leads writing teachers to a consideration that even though there is a rubric for teachers to follow, it does not still assure the unity in scores among graders. Hence, it requires writing teachers to design rubric in a very detailed and precise way; otherwise, teachers should be very experienced in scoring students' papers.

The pedagogical implications can be clearly understood from the findings of the sample analysis and interview stage. Regarding the categories in Rubric 1 such as spelling, grammar, style, and vocabulary words, it can be reliable if the papers are assessed by *PaperRater;* however, the scores generated by the program can be considered only as reference grades. Teachers can rely on the statistics of errors and suggestions assigned by the program to generate appropriate scores. After that, teachers can quickly assess the three extra categories in Rubric 2, which can consume less time. This combination in grading papers can help writing teacher save a great amount of time. In this study, when grading the papers, the two teacher scorers were asked to estimate the time to complete scoring each paper. The result shows that it averagely takes 20 to 25 minutes to fully score a prompt following the criteria in the rubrics. However, it takes less than 1 minute for *PaperRater* to complete grading all the categories in Rubric 1. Hence, writing teachers can take advantages of *PaperRater* to save time and effort.

In short, the findings of this study indicate that the reliability of *PaperRater* is acceptable and that writing teacher can somehow rely on the functions of *PaperRater* as a reference in grading papers. An appropriate combination of traditional and computerized grading methods can generate effectiveness, especially in large classrooms or with a great number of papers. However, this study has only been conducted on low-leveled students; the result might be different in advanced levels, on which it may require further research. One more limitation that can be found is that it would be better if there were more teacher scorers. This can help the author confidently assume the reliability and correlation among the sets of scores. Nevertheless, the findings in this research can significantly suggest a new method for ESL/EFL writing teachers to grade student's writing assignments, especially for those who simultaneously deal with a number of large classrooms.

ACKNOWLEDGEMENTS

I am really grateful to my colleague and students in Dalat University who willingly helped me establish the data for my study. Also, I would like to thank Dr. Victoria Rau and TA. Ann Chang in National Chung Cheng University for their helpful comments on an earlier draft of the paper.

REFERENCES

Attali, Y., & Burstein, J. (2004, June). *Automated essay scoring with e-rater V.2.0.* Paper presented at the Conference of the International Association for Educational Assessment, Philadelphia, PA.

Burstein, J. (2003). The e-rater scoring engine: Automated essay scoring with natural language processing. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 113–121). Mahwah, NJ: Erlbaum.

Chapelle, C. A., and Douglas, D. (2006). *Assessing language through computer technology.* Cambridge: Cambridge university Press. 138 pp.

Cohen, Y., Ben-Simon, A., & Hovav, M. (2003, October). *The effect of specific language features on the complexity of systems for automated essay scoring*. Paper presented at the 29th Annual Conference of the International Association for Educational Assessment, Manchester, UK.

Elliot, S. M., & Mikulas, C. (2004, April). *The impact of MY Access!™ use on student writing performance: A technology overview and four studies*. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Hossein, F. (2012). Principles of Language Assessment. In Christine, C., et al. (Eds.), *The Cambridge Guide to Second Language Assessment.* (pp. 37-46). New York, USA: Cambridge University Press.

Keith, T. Z. (2003). Validity and automated essay scoring systems. In M. D. Shermis, & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective*. (pp. 147–167). Mahwah, NJ: Erlbaum.

Landauer, T. K., Laham, D., & Foltz, P. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 87–112). Mahwah, NJ: Erlbaum

National Commission on Writing in America's Schools and Colleges. (2003). *The neglected "r": The need for a writing revolution*. New York: The College Entrance Examination Board.

Shermis, M. D., Mzumara, H. R., Olson, J., & Harrington S. (2001). On-line grading of student essays: PEG goes on the World Wide Web. *Assessment & Evaluation in Higher Education*, *26*(3), 247–259.

Shin, S. Y. (2012). Web-Based language testing. In Coombe. C., et al. (Eds.), *Second language Assessment* (pp. 274-279). Cambrige University Press.

Warschauer, M., & Grimes, D. (2008). Automated Writing Assessment in the Classroom. *Pedagogies: An International Journal*, 3, 22–36.

About the Author

Nguyen Vi Thong is currently a Ph.D student at Graduate Institute of Linguistics, National Chung Cheng University, Taiwan. He is also the Vice Dean of the Department of Foreign Languages, Duy Tan University, Vietnam.

APPENDICES

Appendix A: Writing Rubric 1

Categories	**A** **90-100** **Excellent**	**B** **80-89** **Very Good**	**C** **70-79** **Satisfactory**	**D** **Below 70** **Unsatisfactory**
Spelling	- correct spelling, punctuation, and capitalization; or hardly found errors.	- a few spelling, punctuation, and capitalization errors.	- Shows a pattern of errors in spelling, punctuation, and capitalization. Could also be a sign of lack of proof-reading.	- continuous errors
Grammar	- correct grammar and syntactic structures; or hardly found errors.	- a few errors found in grammar and syntactic structures.	- Shows a pattern of errors in grammar and syntactic structures. Could also be a sign of lack of proof-reading.	- continuous errors
Word Choice	- appropriate words used. - No/Hardly bad phrases found.	- a few inappropriate words or bad phrases found in the essay.	- inappropriate words or bad phrases often found in the essay.	- no attempt at choosing appropriate words or good phrases.
Style	- good usage of transitional words/phrases.	- transitional words/phrases often found in the	- transitional words/phrases sometimes found in	- transitional words/phrases rarely/hardly

	- various sentence structures used including simple, compound, complex, and mixed types.	essay. - different sentence structures used, but not many.	the essay. - there is an attempt at varying sentence structures, but sometimes leaves some errors.	found in the essay. - There is almost no attempt at varying sentence structures.
Vocabulary words	- considerably attempts to use various academic words.	- academic words can be often found through the essay.	- academic words can be sometimes found through the essay.	- There is almost no attempt at using academic words.

Appendix B: Writing Rubric 2

Categories	**A** **90-100** **Excellent**	**B** **80-89** **Very Good**	**C** **70-79** **Satisfactory**	**D** **Below 70** **Unsatisfactory**
Spelling	- correct spelling, punctuation, and capitalization; or hardly found errors.	- a few spelling, punctuation, and capitalization errors.	- Shows a pattern of errors in spelling, punctuation, and capitalization. Could also be a sign of lack of proof-reading.	- continuous errors
Grammar	- correct grammar and syntactic structures; or hardly found errors.	- a few errors found in grammar and syntactic structures.	- Shows a pattern of errors in grammar and syntactic structures. Could also be a sign of lack of proof-reading.	- continuous errors
Word Choice	- appropriate words used. - No/Hardly bad phrases found.	- a few inappropriate words or bad phrases found in the essay.	- inappropriate words or bad phrases often found in the essay.	- no attempt at choosing appropriate words or good phrases.
Style	- good usage of transitional words/phrases.	- transitional words/phrases often found in the	- transitional words/phrases sometimes found in	- transitional words/phrases rarely/hardly

	- various sentence structures used including simple, compound, complex, and mixed types.	essay. - different sentence structures used, but not many.	the essay. - there is an attempt at varying sentence structures, but sometimes leaves some errors.	found in the essay. - There is almost no attempt at varying sentence structures.
Vocabulary words	- considerably attempts to use various academic words.	- academic words can be often found through the essay.	- academic words can be sometimes found through the essay.	- There is almost no attempt at using academic words.
Main idea	- Clearly presents a main idea and supports it throughout the paper.	- There is a main idea supported throughout most of the paper.	- Vague sense of a main idea, weakly supported throughout the paper.	- No main idea
Organization	- Well-planned and well-thought out. Includes title, introduction, statement of main idea, transitions and conclusion.	- Good overall organization, includes the main organizational tools.	- There is a sense of organization, although some of the organizational tools are used weakly or missing	- No sense of organization
Content	- Exceptionally well-presented and argued; ideas are detailed, well-developed, supported with specific evidence & facts, as well as examples and specific details.	- Well-presented and argued; ideas are detailed, developed and supported with evidence and details, mostly specific.	- Content is sound and solid; ideas are present but not particularly developed or supported; some evidence, but usually of a generalized nature.	- Content is not sound

Appendix C: Interview Questions

No.	Prepared Questions	Follow-up Questions
1	What are the advantages and disadvantages of scoring papers based on the rubrics?	Do you have any suggestion on how to design the rubrics in more scorer-friendly way? Do you often have a detailed rubric for every student's assignment?
2	How important do you think intuition and experience are in scoring writing prompt?	Do you think intuition and experience may cause difference in scores among writing graders? To what extend do you think so? Do you think each teacher's personalities may affect the writing evaluation?
3	How large are you writing classes? How much time do you have to spend on scoring their papers? How do you feel if there is a program that helps you score some parts of the papers? Will you totally believe in the scores generated by the program?	You said that you would not totally rely on the scores assigned by the program, so do you have any suggestion on how the software should be used in writing classrooms?