International Journal of Information Technology and Language Studies
(IJITLS)

Automated versus Human Essay Scoring: A Comparative Study

Rania Zribi and **Chokri Smaoui**
raniazribi@ymail.com; smaoui2002@yahoo.com

Faculty of Letters and Humanities, University of Sfax, Tunisia

Abstract. The purpose of this study was to investigate the validity of automated essay scoring (Paper Rater) of EFL learners' written performances by comparing the average group mean scores assigned by the Paper Rater computer and by human raters. Ten intermediate EFL learners responded to a topic and received scores from both automated and human scoring processes. The SPSS statistical procedure, namely the One-Way Reported-Measures ANOVA, diagnosed the difference between the computerized mean scores and human raters' mean scores. Unlike previous studies, the findings of this study reflected some differences in the scores awarded by both procedures. The average mean scores assigned by the automated essay scoring tool Paper Rater was significantly higher than the human raters' scores of learners' essays. The Paper Rater tool did not seem to correlate well with human raters. Thus, the implications for English teachers revealed that despite its cost-effective nature, the automated scoring system together with human scorers lack the ability to award as reliable scores as humans do. However, the application of computerized scoring system in the educational system plays a key role in improving the learning process. Thanks to its instant feedback, this software may contribute to the improvement of EFL learners' writings.

Keywords: the automated essay scoring (Paper Rater), human scoring, reliability, variation, feedback, writing improvement.

1. Introduction

One of the essential features of human language is its ability to be written. Writers display different language features while conveying their written messages depending on the purpose and the context of performances. Thus, writing is "an act that takes place within a context, that accomplishes a particular purpose, and that is appropriately shaped for its intended audience" (Hamp-Lyons & Kroll, 1997 p.8). As a formal mode of interaction, the writing skill plays an increasingly important role in the community. Indeed, it is a means of communication that permits people to exchange information, shape, and transmit their messages indirectly in the form of words, sentences, paragraphs, or essays. This goes in line with Caswell and Mahler's (2004) view that "writing is the vehicle for communication and a skill mandated in all aspects of life" (p.3).

Due to its dynamic nature in social interaction, special attention should be directed towards integrating the practice of the learners' writing skills in the first, second and foreign language educational system. Academic writing has thus a privileged role in English for Academic Purposes instruction (EAP). The latter refers to "language research and instruction that focuses on the specific communicative needs and practices of particular groups in academic contexts" (Hyland & Hamp-Lyons, 2002 p.2). The development of the writing skill is highly recommended in both L1, L2 and foreign language instruction depending on the learners' different learning interests, needs and objectives in various academic contexts. Hence, the assessment of EFL learners' written productions emerged as an essential issue to test the students' performances, develop their writing abilities, and help them communicate effectively.

Testing written performances is beneficial for learners whose aim is to improve their learning process based on teachers' scores and feedback to tests. By relying on a well-defined rating scale with its welldetermined criteria, teachers give precise evaluative comments on different aspects of language to their test takers' essays to help them recognize their problems in writing a piece of paper and to motivate them to be engaged in a more creative and responsive writing course. In this vein, Brookes and Grundy (2001 p. 2) stressed the importance of feedback by stating that "we pay more attention to writing since we are more aware of what we are doing and consequently we give more emphasis to correctness".

However, a problem may arise concerning the scorers of EFL learners' writing skills. Despite the use of the same rating scale and criteria by different examiners, the measurement of compositions may vary from one rater to another due not only to personal and professional factors related to testers, but also to the subjective nature of essay assessment. In fact, subjective scorers' judgments have an impact not only on scores' validity and test reliability, but also on test takers' writing ability (Bijani, 2010 p.70). Raters' subjectivity was further highlighted by Peterson (2008), who reiterates that evaluating learners' written productions is a subjective process (p.72), resulting in potential scores variation and differences in scoring patterns and reading styles due to raters' factors (Lumley & McNamara, 1995). Moreover, raters may fail to provide precise and timely feedback to their students to enable them to improve their abilities and convey their ideas effectively because the testing process is time-consuming, especially for large classes.

Thus, researchers and educators pay attention to the application of technology in the language testing field. Hence, the automated essay scoring system major role is its ability to assign scores to essays and to offer clear and prompt feedback in just few seconds. To support this view, Bennet and Ben-Simon (2005) claimed that "automated essay scoring has the potential to reduce processing cost; speed up the reporting results, and improve the consistency of grading" (p.3). In order to ensure reliable, valid and fair assessments, raters' potential scores variability should be minimized. Raters' subjective judgments may lead to unfair evaluations. This is why educators, with the emergence of new technological devices in the educational field, put special emphasis on how to use technology in evaluating learners' language proficiency. Hence, more research should be carried out to shed new light on the existing studies of automated essay scoring patterns and strategies.

The purpose of the current research is to investigate the reliability and the usefulness of automated essay scoring (AES) for classroom-based tests by comparing the group mean scores assigned by human raters with marks provided by the automated essay scoring system after testing the same set of essays produced by ten test takers based on a well-defined rating scale. Our major aim is to compare scores obtained from both computerized scoring system and human raters to examine their degree of reliability and consistency in evaluating EFL learners' written skills.

This study sought answers to the following research questions:

1. Is the group mean score awarded by the automated essay scoring (Paper Rater) significantly different from the group mean score provided by human raters in testing EFL test takers' compositions?

2. Is the reliability rate among human raters different from that between human raters and the Paper Rater scoring system?

3. Does computerized Paper Rater tool result in significant improvement of foreign language learners' writing achievement?

2. Review of the literature

Writing performance assessment has become a vital issue for both teachers and researchers in the language educational system as the writing skill has become more appealing in language teaching than before, due to the implementation of the communicative approach in education. Since technology has invaded our life in myriad fields, educators and teachers have looked for effective ways to facilitate their assessment process and ensure reliable, valid, and fair measurements of their students' language skills. Indeed, Researchers in this context, have "struggled with the development of methods able to produce a reliable and valid means of directly assessing writing quality" (Huot, 1990 p.237). In this vein, Williamson at al. (2010) for instance claims that human rating is not the only option, where technology is available everywhere, to mark learners' productions.

Hence, further research should be conducted to assess the usefulness of technological devices; especially automated evaluation systems in measuring learners' writing skills and improving their writing proficiency. Automated Essay scoring (AES) can be defined as the computer technology that assesses and rates learner's written performances. Nowadays, AES systems play a pivotal role in the teaching, learning, and testing fields. In fact, they are used to overcome time, cost, reliability and generalizability problems in the language testing domain. (Dikli, 2006, p.3).

How well does automated essay scoring (AES) correlate with human raters' scoring? A number of researchers and educators tried to give an answer to this question. For instance, Page (1968), the inventor of a computer grading program named Project Essay Grader (PEG), conducted a study in which he compared the scores obtained from his PEG with those received from human raters after rating students' compositions. Based on the multiple R correlation statistical procedure, Page found a positive correlation coefficient of 0.77 between the automated and the human scoring systems.

In support of Page's claim, Nivens-Bower's (2002) comparative study revealed consistent results from computerized scoring, the IntelliMetric program, and human scoring. Both the paired-sample t-test and the Wilcoxon signed rank test reflected not only no significant difference in both the group means and range of score frequencies, but also high correlational coefficient rates in the marks assigned by both rating systems. Hence Nivens-Bower stated that IntelliMetric "produced results consistent with what would be expected of faculty scores" (p.12). Foltz et al.'s (1999) findings corroborate with previous results as their candidates' essay scores received from both human raters and the Intelligent Essay Assessor (IEA) computer program mirrored a significant inter-rater correlation rate.

In the same context, Wahlen et al. (2020) conducted a comparative study in which they examined the scores assigned by the automated scoring system ESCRITO and human raters to learners' responses to open-ended tasks. Authors perceived a significant agreement between automated and human ratings. They also ascertain that the automated scoring system measures the same construct as the human rating. This assumption is based on the fact that the rating outcomes of both computers and humans were convergent (p.8). In this respect, a number of researchers highlighted the usefulness of implementing computer-assisted grading applications as they "did not negatively affect student attitudes concerning the helpfulness of their feedback, their satisfaction with the speed with which they received their feedback, or their satisfaction with the method by which they received feedback" (Anglin, Anglin, Schumann, & Kaliski, 2008, p. 51). In agreement with this view, Cohen et al. (2018) were in favor of applying automated evaluation as a valuable complement in the scoring process. Their automated essay system (AES) was able to analyze lexical and semantic features rather than discourse features (p.15). It did not consider the same concept of what constitutes a good writing as that held by human markers. Hence, Cohen et al. (2018) claim that "AES scoring is not a perfect substitution for human scoring, but can be a useful complement to it" (p.16).

While some researchers were in favor of applying the automated essay scoring programs in testing EFL learners' written performances, other linguists and educators criticized the use of computerized scoring systems in scoring the test takers' writing skills as they found a negative relationship between the scores received from both types of assessment. In this respect, in a study conducted by Wang and Brown (2008), writing productions were collected from the test takers' responses to the Writer Placer Plus test. In fact, the writers reported a low score correlation between trained human graders and IntelliMetric computer program. Thus, the use of the Spearman Rank Correlation Coefficient test revealed no significant correlation between the overall holistic automated marks and holistic human marks (p.319).

Similarly, Huang (2014 p.160) came up with the conclusion that automated and human scores are different and are weakly related. He explained his argument by stating that the automated essay scoring program tended to give higher marks than human markers. To confirm the low reliability of the automated scoring system, McCurry (2010) applied two different machines scoring writing software (MSW) to assess open writing task responses. The automated scores were compared with human scores awarded to the same writing test showing a low-reliability rate. In his paper, McCurry concluded that 'automated essay scoring' (AES) did not grade the broad and open writing task responses as reliably as human markers (p.127). These results are consonant with Anson's (2003) claim that "it is nearly impossible for AES tools to imitate the human assessment process, which involves 'multiple subjectivities' and 'sophisticated intellectual operations" (p.236).

In the same vein, Shermis (2015) compared the scoring performance of the Automated Student Assessment System (ASAS) with its eight rating engines to that of trained human raters. The results of the statistical analysis revealed some differences between the two rating modes. Human raters reached a complete level of agreement (kw= 0.89). However, the level of agreement between the eight automated rating engines was low (kw= 0.72). Hence, Shermis (2015) elucidates that the automated scoring system did not achieve the same degree of agreement as to the human raters. This can be due to the complexity of writers' correct response options. In fact, human writers can perform different variations of words, phrases, and sentences that are easily recognized by human graders but can be easily overlooked by the automated scoring system. This may lead to different systematic biases that affect reliability, validity, and fairness (p.62).

3. Methodology

3.1 Research design overview

The current study uses a quantitative causal-comparative study design. We focus on two different rater groups, automated essay scoring (AES) and human raters, to diagnose the way human raters (higher education English teachers) and automated essay scoring program (in this case the Paper Rater system) score EFL test takers' written performances after responding to a one-hour writing test based on a well-defined rating rubric comprising six criteria. The independent variable of this study is the manner of rating through which a comparison is held between computer and human scoring results whereas the dependent variables are the holistic group mean scores assigned by both Paper Rater computer software and human raters.

A comparative pattern was useful in this study in order to extract the differences and similarities in the scores assigned by raters and the automated scoring system to EFL test takers' writing samples based on the same holistic rating scale with its well-defined criteria. To advocate the efficiency of the comparative design in analyzing the study outcomes in the language testing field, Collier (1993) argued that "comparison is a fundamental tool of analysis. It sharpens our power of description, and plays a central role in concept-formation by bringing into focus suggestive similarities and contrasts among cases" (p.105).

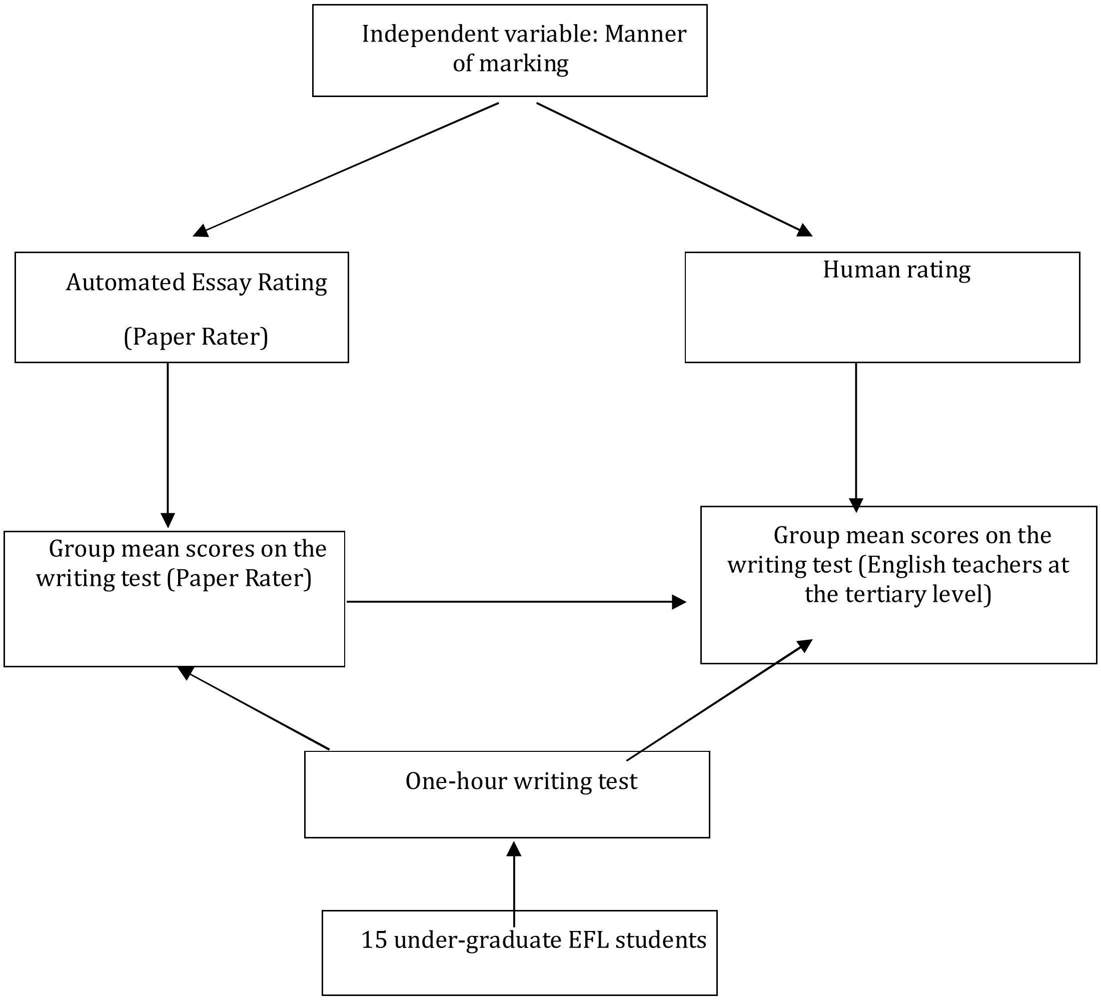
Independent variable: Manner of marking
Automated Essay Rating
(Paper Rater)
Human rating
Group mean scores on the writing test (Paper Rater)
Group mean scores on the writing test (English teachers at the tertiary level)
One-hour writing test
15 under-graduate EFL students

Figure 1. The comparative design model

3.2 Population

A sample of fifteen intermediate foreign language learners, representing a mixed population of males and females, were selected randomly. The target students were enrolled in the Faculty of Letters and Humanities of Sfax, specifically in the English department, to pursue their English language studies for three years. Participants in this study consisted not only of EFL learners but also of English teachers serving as raters whose major role was to assess their test takers' written skills.

3.3 Procedures

To gather data, the EFL examinees responded to a one-hour writing test. Then, their performances were marked holistically by relying on the online automated scoring software the Paper Rater. This computer program instantly analyzed each essay by automatically applying a six-point rating scale, whose focus was directed mainly towards three different aspects of language, namely coherence and connectivity, phrase structure and formation, and vocabulary and word expression. After applying the machine scoring procedure, we contacted ten English teachers at the tertiary level to grade the same set of EFL compositions.

To meet the study's aims, human raters utilized the Magoosh Essay Rubric, which is designed to evaluate the General Management Admission Test for business-examination purposes. The latter focused on five different language criteria to measure the quality of essays. However, following the objectives of our study, some modifications were made in this rubric. We eliminated two criteria from the scale, quality of idea and summary for the sake of obtaining feedback from both automated and human scoring tasks on the same aspects of language. In the adjusted version of the original rubric, only three language categories, namely organization, writing style, and grammar and usage are taken into account. Thus, each criterion has six levels ranging from "lack of proficiency" to "native-like proficiency". Before they started their assessment task, English teachers received some guidelines about how to use the rating scale appropriately during their participation in a twenty-minute training session.

Based on the rating scale, EFL compositions were judged and graded by the two types of the scoring groups. Due to the requirements of the third research question, the fifteen test takers were divided into two groups in the final stage. The first group was exposed to the Paper Rater essay scoring feedback while the second group was exposed to the comments and evaluation of the ten human raters. After their exposition to the evaluation of both rater groups, the fifteen test takers sat for a second writing test to write another piece of paper by taking into account the writing mistakes and weaknesses that they were told about. The final step of the assessment consists of gathering both automated and human scores and entering them in the SPSS database to extract their group mean scores, which enables us to seek any possible correlation between both scoring groups and to compare them by taking into consideration their degree of reliability rate.

Three statistical procedures were run separately in the current study to compare the holistic scores assigned by human raters and those awarded by the automated essay scoring system. Based on the statistical SPSS program, we utilized the one-way repeated measures ANOVA test together with descriptive statistics, namely the mean scores and the standard deviation and the coefficients of variation, to extract the points of similarities and differences between both rater groups overall mean scores. ANOVA's main aim is to examine potential scores variation in both methods of grading EFL compositions. In addition, we resorted to the Intra-Class Correlation Coefficient Test to diagnose the inter-rater reliability rate of human judges and the agreement rate in the marks given by both human raters and the computer essay scoring program. Finally, the Independent Sample T-Test was applied to examine the effect of the automated essay scoring tool on improving student essay writing.

4. Findings and discussion

4.1 Variance

Both the descriptive statistics and the one-way repeated measures ANOVA assess the null hypothesis that there is no statistically significant difference between the group mean score assigned by Paper Rater and the group mean score assigned by human raters on the one-hour writing test. Table 1 shows and compares the mean average scores assigned by the two rating methods, Paper Rater scoring system and human raters, for the fifteen foreign language learners' written responses to a one hour writing test. The automated essay scoring system mean scores were higher than those of the human rater group (Group means of: 4.4420 vs. 3.3800). Thus, the mean difference of both groups was estimated at about 1.062 to reflect a variation in the overall grading of test takers' written performances.

	Means	Standard Deviation	Coefficient of Variation
Computer	4.4420	.76928	17.31
Human	3.3800	.70679	20.910

Table 1: Descriptive statistics for both Human raters and Paper Rater holistic scores

The standard deviation was also applied to measure the degree of variability among the holistic rating methods scores. Hence, the marks awarded by human raters (0.70) were more homogenous and less spread out than those given by the computerized scoring system (0.76). Thus, we can deduce a higher agreement among the ten human rater scores than among computer program scores provided to the same set of candidates' essays. A significant variation rate can be extracted from the coefficients of variation statistical procedure. A high rate of difference (3.70) between the two scoring methods reflects a low uniformity degree of human and automated scores.

As table 2 illustrates, this variation in the mean scores assigned by the two rating methods is further justified by the one-way Repeated Measures ANOVA test results. Based on the F ratio (F > 1 (F=12.128)), the independent variable, the manner of grading (human vs. automated scoring) had an effect on the scores awarded to the EFL learners' compositions. The large F value indicated that the difference between group mean scores was greater than it would be expected by chance or error alone. According to Cohen (1988), both the significance P value (0.07 <0.5) and the effect size value ŋ2=0.57 reflected a statistically significant large effect of the manner of rating on the marks given by each method. Hence, the null hypothesis is rejected. In short, the manner of rating learners' writing skills had a significant effect on the holistic mean scores due to the variation in measuring the ten samples.

Source	F	Wilks's A	Sig	ŋ2
Grading methods (Paper R vs. Humans)	12.128	.426	.007	.574
Error df	9.00			

Table 2: Results of the One-Way Repeated-Measures ANOVA

The significant variation in the scores awarded by human raters and Paper Rater to the ten EFL essays was supported by Wang and Brown's (2007) study in which the IntelliMetric computer program gave higher scores to essays than did human raters. The automated scoring tendency to provide higher scores than did the human scorers may cause a problem especially in high-stakes assessments and for placement purposes. Thus, by obtaining a high mark, the student may be placed at a level that did not suit his learning capacities and led him to face some language difficulties. This view is further supported by Huot's (2002 p.148) claim that "a valid testing tool should be able to reflect whether learners are ready for a specific level of instruction".

4.2 Inter-Rater Reliability

To address the second research question, the Intra-Class Correlation Coefficient test (ICC), using Two-Way Mixed-Effect Model in SPSS, was conducted. Apart from detecting the degree of variance in the group mean scores, our current study aims at examining the degree of inter-rater reliability among the eight human judges while scoring ten essays holistically. Then, the same test was used to diagnose the agreement rate by comparing the scores graded by both humans and computer scoring procedures. In the Two-Way Mixed-Effect Model, the independent variable, namely the grading methods whether human raters or Paper Rater, is the fixed effect while the test takers' compositions are the random effect.

The results of the ICC showed on the one hand a value of .77, an adequate agreement rate in the scores assigned by the eight human markers. This leads to raters' consistency in the measurement of learners' productions (as presented in table3). The latter is further explained by p value of the test (inferior to 0.05), which reflected a significant correlation between human scores. On the other hand, the value of .14 mirrored a poor agreement rate between human scorers and the computerized scoring software (McGraw & Wong 1996).

	Intra-Class Correlation	95% confidence interval		F test with true value 0	
		Lowest bound	Upper bound	value	Sig
Human raters	.770	.458	.932	5.629	.000
Human raters vs. computer scoring program	.141	.-443	.674	1.347	.332

Table 3: Intra-Class Correlation Coefficient (among human raters) and (between human raters and computer scoring

These results are compatible with Wang and Brown's (2008 p.21) correlation study. Human judges are more consistent with each other in testing essays while Paper Rater marks are less consistent with human judges. This difference in the reliability rate can be due to the human subjective nature of testing essays as opposed to the objective machine scoring system. In the same vein, Pilliner (1968) distinguished between subjective and objective rating methods. He argued that:

If the examiner has to exercise judgment; if he has to decide whether the answer is adequate or inadequate; if he has to choose between awarding it a high or low mark; then the marking process is 'subjective'. If, on the other hand, (…) he is reduced for the purpose of marking, to the status of a machine; then the marking process is 'objective' (p. 21).

4.3 Automated scoring system usefulness

The Independent Sample T-Test was applied from the statistical SPSS program to test our second null hypothesis that automated essay scoring system (Paper Rater) does not result in significant improvement of learners' writing achievement in the second writing test. This hypothesis was rejected due to the following t-test results. As the two-tailed P value (0.005) is less than 0.05, a statistically significant difference in the mean scores assigned by both rating groups was perceived in the fourth table. The mean difference scores rate of both scoring methods, estimated of about 1.062 showed a significant scores variation (Human means: 3.3800, automated means: 4.4420) in the first test.

It is in the second writing test that the mean scores of the computerized scoring system increased in a significant way, which was not the case for the human scoring mean scores. In fact, Paper Rater program reached a mean score rate of 6.5432 in the second test whereas the human raters mean scores decreased to 2.4220 rate in the same test (table 5). Hence, we can conclude that the computer program plays a major role in improving the learners' written abilities as the same test takers responded better to the second writing test after they received automated feedback rather than human comments. They wrote better compositions as they responded positively to the Paper Rater evaluation and they took into account the computerized testing.

	Leven'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 variance assumed	.076	.785	3.215	18	.005	1.06200	.33036	.36795	1.75605
Equal variances not assumed			3.215	18	.005	1.06200	.33036	.36795	1.75605

Table 4: Independent Samples T-Test for automated and human scoring essay scoring procedures.

	Human essay scoring	**Computer essay scoring**
Writing test 1	**3.3800**	**4.4420**
Writing test 2	**2.4220**	**6.5432**

Table 5: Group Mean scores for human and automated essay scoring in the first and the second writing test.

The statistical results indicated that the computer scoring essay program is beneficial in the educational system as it assists students to ameliorate their writing abilities. Apart from its ability to improve EFL learners' writing proficiency, the Paper Rater is cost-effective software, which can reduce time-consuming problems and speed up the evaluation task by providing immediate assessment of each learner's performance in each aspect of language based on the well-defined rating scale.

5. Conclusion

The present study has investigated whether the assessments of fifteen learners' written performances in two different writing tests situations differed among human raters and the computerized essay scoring system (Paper Rater) in the EFL context in Tunisia. The major findings of the current study can direct our attention to different implications. The application of technology, like computers, inside classrooms is of great importance as it assists teachers not only in their teaching task but also in their assessment process. Due to its instant evaluation, automated scoring systems may help teachers to gain time in giving clear feedback and may thus improve the test takers' writing abilities. Hence, it is recommended that e-rating essay software in general and Paper Rater in particular could be employed as an educational tool useful for the improvement of the educational system and the learners' language abilities. The teachers' awareness and use of new useful technological procedures to teach the writing skills seem to be important as it permits students to be creative while producing an essay using the target language.

However, we should not lose sight of the fact that automated rating cannot cater for intricacies that only human intervention can watch. This was shown, for example, through the higher agreements that we noted between human raters. By way of illustration, there are subtleties of style and usage that automated rating can hardly detect. This is actually where the human factor comes in strongly. This notwithstanding, there is still a place for the type of automated rating we described in this article. When it comes to practicality, for instance, this instrument can be used to help teachers in rating some tasks, particularly when the stakes of the test are not too high.

References

Anglin, L., Anglin, K., Schumann, P. L., & Kaliski, J. A. (2008). Improving the Efficiency and Effectiveness of Grading Through the Use of Computer-Assisted Grading Rubrics. *Decision Sciences Journal of Innovative Education*, vol.6, (1), pp.51-73.

Anson, C. M. (2003). Responding to and assessing student writing: The uses and limits of technology. In Takayoshi, P. & Huot, B. (Eds.), *Teaching writing with computers: An introduction* (pp. 234–245). New York: Houghton Mifflin Company.

Bennett, R. E. & Ben-Simon, A. (2005). Toward theoretically meaningful automated essay scoring. *Journal of Technology, Learning, and Assessment,* vol.6, (1), pp.1-47.

Bijani, H. (2010). Raters' perception and expertise in evaluating second language compositions. *The Journal of Applied Linguistics*, vol.3, (2), pp.69–89.

Brookes, A., & Grundy, P. (2001*). Beginning to write*. (3rd Ed.) Cambridge: Cambridge University Press.

Caswell, R & Mahler, B. (2004). *Strategies for Teaching Writing.* United States of America: ASCD (Association for Supervision and Curriculum Development).

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd Ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, Y., Levi, E., & Ben-Simon, A. (2018). Validating human and automated scoring of essays against "True" scores. *Applied Measurement in Education*, vol.31, (3), pp. 241-250.

Collier, D. (1993). The comparative method. In A. W. Finifter (Ed.). *Political Science: The State of the Discipline* 2. Washington, D. C. American Political Science Association.

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, vol. 5, (1), pp.1-36.

Foltz, P.W., Laham, D. & Landauer, T.K. (1999). Automated Essay Scoring: Applications to Educational Technology. In Collis, B. & Oliver, R. (Eds.), Proceedings of EdMedia: World Conference on Educational Media and Technology (pp. 939-944). Association for the Advancement of Computing in Education (AACE).

Hamp-Lyons, L. & Kroll, B. (1997). *TOEFL 2000-writing: composition, community, and assessment*. Educational Testing Service.

Huang, S. J. (2014). Automated versus human scoring: A case study in an EFL context. *Electronic Journal of Foreign Language Teaching*, 11(1), pp.149-164.

Huot, B. (1990). Reliability, validity, and holistic scoring: What we know and what we need to know. *College Composition and Communication*, vol. 41, (2), pp. 201-213.

Huot, B. (2002). (Re) *Articulating writing assessment for teaching and learning*. Logan: Utah State University Press.

Hyland, K., & Hamp-Lyons, L. (2002). EAP: Issues and directions. *Journal of English for academic purposes*, vol. 1, (1), pp. 1-12.

Lumley, T., & McNamara, T., F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing,* vol. 12, pp. 54-71.

McCurry, D. (2010). Can machine scoring deal with broad and open writing tests as well as human readers? *Assessing Writing*, vol.15, 2, pp.118-129.

McGraw, K.O., & Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods,* 1, (1), pp. 30–46.

Nivens-Bower, C. (2002). Faculty-Write Placer Plus score comparisons. In Vantage Learning, Establishing Write Placer Validity: A summary of studies (p. 12). (RB-781). Yardley, PA: Author.

Norusis, M. J. (2004). SPSS 12.0 guide to data analysis. Upper Saddle River, NJ: Prentice Hall.

Page, E. B. (1968). The use of the computer in analyzing student essays. *International Review of Education*, vol.14, 2, pp.210–225.

Peterson, S., S. (2008). *Writing across the curriculum*: All teachers teach writing. Portage & Main Press.

Pilliner, A, (1968). Subjective and objective testing. In Davies, A. (Ed.), *Language Testing Symposium*: A Psychological Approach. Oxford University Press. London.

Shermis, M. D. (2015). Contrasting state-of-the-art in the machine scoring of short-form constructed responses. *Educ. Assess.* Vol.20, pp. 46–65.

Wahlen, A., Kuhn, C., Zlatkin-Troitschanskaia, O., Gold, C., Zesch, T., & Horbach, A. (2020). Automated Scoring of Teachers' Pedagogical Content Knowledge-A Comparison between Human and Machine Scoring. *Frontiers in Education*, Vol. 5, p. 1-10.

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

Wang, J., & Brown, M.S. (2007). Automated essay scoring versus human scoring: A comparative study. *Journal of Technology, Learning, and Assessment,* vol.6, 2,pp. 1-29.

Williamson, D. M., Bennett, R. E., Lazer, S., Bernstein, J., Foltz, P. W., Landauer, T. K., .Sweeney, K. (2010). *Automated scoring for the assessment of common core standards*. White Paper