The Use of an Automated Writing Evaluation System for Summative Assessment in an EFL Context: The Relationship Between Automated System Scores and Human Raters' Scores

Hilal Yıldız[1], Safiye İpek Kuru Gönen[2]

[1]School of Foreign Languages, Antalya Bilim University, Antalya, Turkey

[2]Education Faculty, ELT Department, Anadolu University, Eskişehir, Turkey

Abstract

EFL students are constantly involved in evaluation of their writing skills to pass language tests and improve their writing in L2. Scoring individual essays and providing feedback may become an onerous task for teachers and they may need assistance in writing evaluation. As an outcome of recent technological advancements, Automated Writing Evaluation (AWE) systems have been utilized to assist teachers in scoring written essays and providing feedback. This study aims at investigating the relationship between teachers' holistic scores and AWE holistic scores given to the same student essays. The participants of the study were seven (n=8) EFL instructors working at a university in Turkey. In line with the study's aim, five students' essays from each instructor's class (n=35) were scored independently by the instructors, the AWE system, and another instructor who served as an external rater. A Spearman rank test was carried out and the findings revealed a positive and statistically significant correlation both between AWE scores and the instructors' and between the scores of the instructors and the external rater. The findings indicate that the correlation between a human-rater and an AWE system can be compatible with the correlation between two human-raters. Moreover, the teachers noted that there were similarities between their assessment and the AWE assessment in terms of the similarity of the scores, the score intervals, and the rationales for the scores in each interval. All in all, the findings may contribute to a better understanding of using these systems in classroom-based writing assessment in EFL contexts.

Keywords: automated writing evaluation; automated essay scoring; computer-assisted writing evaluation; writing assessment in EFL classes; writing in L2

1. Introduction

The ability to write is an important talent for academic achievement since topic knowledge is typically assessed through writing, particularly in higher education, and students are continuously involved in academic writing (Cai, 2013; French, 2020). Since writing is conceived as the most complex talent among other language skills (Lerner, 1996), acquiring writing skills in an English as a foreign language (EFL) setting is assumed to be more difficult than in an L1 context (Silva, 1993). The main reasons of the difficulty faced by students are the interlingual errors which arise as a result of students' tendency to transfer linguistic patterns and forms from their original language, and intralingual errors which occur as a result of students' insufficient understanding of the target language itself (Richards, 1974). To assist students in becoming more proficient L2 writers, overcoming the obstacles of learning the language, and gaining competence in L2 writing, learners' writing abilities are assessed regularly in EFL context.

1.1. Assessment of writing in EFL context

Assessment might be defined as anything that students and teachers undertake in the classroom that results in decision-making and reflection (Pearson, 1998). Assessment is an essential component of the teaching and learning process since it identifies learners' strengths and limitations. Although the terms 'assessment' and 'evaluation' are often used interchangeably in the educational context, evaluation refers to the act of assessing performance in order to evaluate the quality or value of performances is referred to as evaluation (Scriven, 1991). Similar to assessment, evaluation provides teachers with information to help them customize their lessons. Learners, like instructors, gain from assessment and evaluation by recognizing their own shortcomings and concentrating on the concerns raised by the assessment in order to better themselves.

The two most common types of assessment utilized in EFL writing are formative and summative assessment. While the primary goal of the former is to offer proof of achievement to students, their parents, or institutions and to give evidence for students' comprehension and learning through scores (Earl & Katz, 2006), the latter refers to giving feedback to bridge the gap between present and desired levels of learning. The information obtained from the assessment is used to develop and adjust future learning activities (Black, Harrison, Lee, Marshall & William, 2004).

In the EFL writing context, many challenges are faced during both summative and formative assessment processes. With respect to formative assessment, providing continuous formative feedback is central to the development of students' L2 writing abilities. Nonetheless, providing continuous, quality feedback can be impractical and time-consuming for the teachers (Long, 2013). Furthermore, providing the classroom-based formative assessment is a teacher-centered process and does not allow learners to be autonomous. It is also worth the mention that personal factors such as lack of knowledge and training and physical and mental fatigue may affect teacher-based formative assessment negatively.

In terms of summative assessment, the information obtained via assessment acts as an indicator and plays a part in decision-making processes, the validity and reliability of an assessment are critical. As a result, teachers have an extra obligation to ensure the correctness and fairness of the assessment, as they are often the ones who gather evidence and report what has been taught (Gardner, 2012). Moreover, it is argued that the raters' personal identity and attitude may result in grade discrepancies, or the social relationships with the students may affect the rater's opinion of the quality of the written content due to a phenomenon known as the "halo effect" (Schaefer, 2008). Aside from inter-rater disagreements, the fluctuation of ratings assigned by the same rater for the same written content may also be inconsistent (McNamara, 2000). Finally, there are time and place constraints in summative assessment due to large number of students in classes, teachers' other duties and the dependency on teacher-based assessment. In order to overcome these challenges reported here, computer-mediated assessment such as AWE systems may be offered as an assistant means of evaluating writing and responding to learners' work, and there is a widespread interest in research examining the possibilities of AWE tools in EFL classes.

1.2. The use of AWE tools in summative assessment of L2 writing

For the sake of reliability of summative assessment, it is underlined that rater training is essential for any rating scale to provide objective scores (McNamara, 2000). It is also advised that written texts can be assessed by more than one assessor so that the final result can more accurately reflect the quality of writing. However, due to expected standards of summative assessment and the aforementioned challenges, it is argued that AWE tools can produce reliable scores compatible with teachers' scores while reducing teachers' burden (Cotos, 2010; Lee, 2017).

AWE tools, which are used to provide a formative and summative assessment of writing skills, can generate scores using machine learning, natural language processing (NLP), and artificial intelligence (AI). Despite the fact that AWE software may use both analytical and holistic ratings, most web-based AWE tools generate scores based on holistic rubrics. Prior research in the use of AWE tools in summative assessment of L2 points to the likelihood of achieving a significantly high positive correlation between human raters and automated raters (Attali & Burstein, 2006; Burstein & Chodorow, 1999; Shermis, Koch, Page, Keith &Harrington, 2002). When Burstein and Chodorow (1999) compared the expert raters' scores given to non-native English speakers' essays written for Test of Written English (TWE) to those of the e-rater system, it was discovered that the correlation coefficients between e-rater scores and human raters were very high and positive (r=0.73), almost as high as the inter-rater reliability between two human raters (r=0.75). Likewise, Powers et al. (2002) compared human and e-rater scores and concluded that the scores given by humans and e-rater were nearly identical because the agreement between human raters on different prompts was between 68 and 94 percent, and the agreement between human raters and E-rater was between 48 and 93 percent. Furthermore, the research found that agreement between AWE scorers and human raters might be more trustworthy than the agreement between two human raters (Shermis et al., 2002; El Ebyary &Windeatt, 2010, Ramineni, 2013). Shermis et al. (2002), for example, evaluated the scores provided to 300 online placement essays by both an AWE tool and professional raters. According to the study, the agreement between the PEG automated writing tool and six experts was strongly and favorably connected (r=83), but the correlation between human raters was significantly lower (r=51).

Research done in classroom settings in native English-speaking (NES) environments yielded conflicting results. Wilson and Roscoe (2020), for example, claimed that the correlation coefficient between human raters and AWE systems may not be as strong as reported in earlier studies since raters' backgrounds and the study context might influence the results. According to the findings of research done by Wilson and Roscoe (2020) at a middle school with native speakers, the correlation coefficient between human raters was very high (r= 0.98), but the correlation between the PEG program and human raters was lower (r= 0.62). On the other hand, Attali and Burstein (2006) discovered a nearly perfect agreement (r= 0.97) between an AWE tool and human raters in research and indicated that AWE systems, such as e-rater, can provide improved standardization of ratings, therefore validating the AWE scores.

Despite an increase in the quantity of literature on ESL and EFL settings over the last decade, the generalizability of published research on AWE validity in EFL and ESL contexts is difficult due to the small number of studies. Despite the fact that research in EFL contexts is uncommon, El Ebyary and Windeatt conducted a promising study (2010) and revealed that AWE scores were substantially and positively associated (r=.83) with one of the raters and moderately linked (r=.59) with the second-rater in this study, whereas the connection between two human raters was lower and moderate (r=.45). To put it in a nutshell, although the growing body of literature has reported favorable findings there is still room for research conducted in EFL contexts in terms of usability and practicality of AWE systems in L2 writing assessment. What is more, teachers' opinions on the use of AWE tools for summative assessment can also help to understand whether these systems are compatible with teacher scoring. Therefore, in an attempt to shed light on the use of AWE tools for summative assessment purposes in a Turkish EFL context, the following question was posed:

- Is there a relationship between AWE holistic scores and instructors' holistic scores given to students' essays?

2. Method

The current study was carried out at the School of Foreign Languages of a foundation university in Turkey. The participants were eight (n=8) non-native EFL instructors, with one serving only as an external rater. In the context of the study, students were required to write essays as part of the writing assessment, and teachers graded these essays and provided feedback throughout the term. All of the seven instructors who took part in the study graded their own classroom papers. Following that, five essays from each class were randomly chosen from the papers gathered, and these essays were analyzed and reviewed by each teacher to be compared with AWE scores. SPSS (20) software was used to detect the correlation between the AWE scores and the instructors' scores. The correlation between the instructors' scores and the AWE scores of the same essays were compared with the correlation between each of the seven instructors and the additional eighth instructor, who was referred to as the external rater in this study. Because the essays are reviewed by two instructors in the EFL higher education setting to validate the results, the eight instructors only engaged in this study to assist in identifying how relevant and suitable computer-generated scores can be in the specific context.

After collecting students' data through their essays, the instructors were informed about the system regarding its functions and features. Furthermore, teachers were given access to both the instructor account and the student accounts, to which the papers gathered from them were posted. They were also briefed about the study's main findings and asked to take notes on their observations regarding their summative assessment of papers and the systems’ summative assessment and scores. The teachers' views on the relationship between their scores and the AWE scores were gathered through semi-structured interviews. Since the AWE scores did not fulfill the normality requirement ($p< .05$), a non-parametric Spearman analysis was used to evaluate if there was a statistically significant connection between instructor scores and AWE scores as well as between instructor scores and external rater scores. Teachers’ opinions gathered through semi-structured interviews were used to support the quantitative findings regarding the relationship between teacher scores and AWE scores.

3. Results and Discussion

Once all the seven instructors scored the essays of the students in their classes, the external rater assessed the essays randomly selected by the other teachers. Furthermore, All of the essays (n=35) were transcribed and submitted to the AWE system in order to get computer-generated scores. The essays were graded using holistic rubrics by both the teachers and the AWE system. The instructors utilized a holistic rubric used in the research context, and the AWE system graded the essays using a six-point holistic rubric. Prior to the correlational analysis, the ratings assigned by the instructors and the AWE instrument were normalized to range from 0 to 100 points. To examine the relationship among the instructors’ scores, the external rater’s scores and the AWE scores a nonparametric Spearman rank analysis was carried out. Table 3.1 below summarizes the findings of the analysis.

Table 3.1. *The correlation among the scores of instructors, the AWE system, and the human-rater*

Variables		**Instructors' scores**	**AWE scores**	**External rater's scores**
Instructors' scores	***r***	1	.862*	804*
	p		.000	.000
	N	35	35	35
	$\bar{x}$	70,64		
AWE scores	***r***	.862*	1	.876*
	p	.000		.000
	N	35	35	35
	$\bar{x}$		59,76	
External rater's scores	***r***	.804*	.876*	1
	p	.000	.000	
	N	35	35	35
	$\bar{x}$			66,28

*p<.05

When the table is reviewed, it can be seen that the average score for teachers was x=70,64, the average AWE score was x=59,76, and the average external rater score was x=66,28. The correlation coefficient between the mean scores of the instructors and the mean scores of the AWE was r = 0.862, and the determination coefficient was r^2 = 0.743. That is, as the AWE scores increased, the teachers' scores increased as well, and when the scores of one variable decreased, the scores of the other variable decreased as well. Furthermore, the correlation coefficient between instructors' and external rater's scores was r = 0.804 and the determination coefficient was r^2 = 0.646 indicating that instructors' and external rater scores increased and decreased simultaneously. As a result, the correlation between instructors' scores and AWE scores as well as the relationship between instructors' scores and external rater scores was statistically significant (p < .05). Furthermore, when the interval levels mentioned in Alpar's (2010) correlation coefficient interpretation criteria were considered, it was discovered that there was a strong, positive, and significant correlation both between the instructors' scores and the AWE scores (r =. 862, p< .05) and between the scores of the instructors and the external rater (r =.804, p< .05). A nonparametric Spearman analysis was also utilized to evaluate the agreement between each instructor's individual scores and the AWE score, as well as the agreement between the instructors' scores and the external-rater scores. Table 3.2 summarizes the findings of the analysis.

Table 3.2. *The correlations between AWE scores, external rater scores, and each instructor's scores*

Variables		**Instructors' scores**	**AWE scores**	**External rater's scores**
Teacher 1	***r***	1	.825	.649
	p		.086	.236
	n	5	5	5
	$\bar{x}$	73	66,4	71
Teacher 2	***r***	1	,975*	1*
	p		,005	.000
	n	5	5	5
	$\bar{x}$	67,5	59,76	69
Teacher 3	***r***	1	.949*	.872
	p		.014	.054
	n	5	5	5
	$\bar{x}$	69	63,08	65
Teacher 4	***r***	1	.975*	1*
	p		.005	.000
	n	5	5	5
	$\bar{x}$	71	56	66
Teacher 5	***r***	1	.740	865
	p		.152	058
	n	5	5	5
	$\bar{x}$	74	56	63
Teacher 6	***r***	1	.866	.564
	p		.058	.322
	n	5	5	5
	$\bar{x}$	73	59,76	64
Teacher 7	***r***	1	.918*	.803
	p		.028	.102
	n	5	5	5
	$\bar{x}$	67	56,44	65

*p<.05

Although the AWE scores appeared to be lower than the instructors' and external rater scores, a high, positive, and significant correlation was discovered between the AWE scores and the second teacher (r =. 975, p <.05), the third teacher (r =. 949, p <.05), the fourth teacher (r =. 975, p <.05), and the seventh teacher (r =. 918, p <.05). This demonstrates that both external rater scores and AWE scores may correlate well with other raters (teachers) and that the correlation between AWE scores and other raters can be stronger in some situations. Table 3.2 also shows that, in general, when the correlation between the AWE scores and the instructors' scores was higher, the correlation between the instructors' scores and the external rater scores was higher as well.

In addition to the quantitative analysis, instructors were asked to check the student AWE accounts and review the scores given by the AWE system to the same essays they scored. In addition, the teachers were asked to take reflective notes about their views on AWE scoring. During the interviews, the teachers were asked what they thought about AWE scores and how they compared the scores to their own. It was revealed that six instructors found AWE scores comparable to their own, while only one instructor found AWE scores lower than her own. The instructors also added that despite the broad range, the similar scores they gave were at the same intervals of the rubrics used on the system as well, and they agreed with the comments/feedback provided by the system justifying the scores. As a result, most of the teachers were pleased with the scores and agreed that the AWE scores were compatible with their scores.

The findings on the correlations between instructors' scores, AWE scores, and external rater scores correlate favorably with the majority of research in the literature. In line with the previous studies (Attali & Burstein, 2006; Shermis et al., 2002), it was found that there is a positive and significant correlation between human rater scores and AWE scores. This put forward that AWE scores might be compatible with human raters' scores and AWE tools can be utilized in classroom-based summative assessment. Moreover, the findings substantiate previous findings (Burstein & Chodorow, 1999; El Ebyary & Windeatt, 2010; Ramineni, 2013) showing that in some cases, the correlation between AWE scores and other raters can be stronger than the correlation between two human raters. It is worth mentioning that due to heavy workload, fatigue and personal differences human raters' scores might vary and the discrepancy between scores might be an issue. In such cases, the use of AWE may help teachers to standardize their scores and resolve the discrepancy between scores. All in all, it can be inferred that these systems continue to develop with the development of technology and considering the workload of the teachers, the compatibility of the scores these systems can be used to create more effective learning environments for L2 writing assessment while reducing the workload of the teachers.

4. Conclusion

Using AWE as a secondary or a supplementary evaluator is becoming increasingly prevalent particularly in high-stakes assessment environments. Research, in general, demonstrates that AWE scores correlate positively and significantly with human raters (Attali & Burstein, 2006, Burstein & Chodorow, 1999, Ramineni et al., 2012, Shermis et al., 2002). Given that two raters are generally involved in the evaluation of writing performance and that the mean scores of these two raters are taken into account in EFL context, AWE can be employed as a second-rater or to settle differences and resolve discrepancies between human raters. In this regard, AWE looks to be a viable and useful tool for use in EFL/ESL situations to improve the scoring process, save time and energy, and help teachers in L2 writing assessment. Especially in EFL contexts, such as the one in this study, where instructors have multiple writing classes with a large number of students, AWE tools may lessen the burden of the instructors and allow them to devote more time to improving teaching. Finally, the study's findings may lead to the conclusion that AWE scoring could be used in higher education settings for summative assessment and they may also be used to inform students about their performance and to urge them to assess themselves by using such systems. This study reported here is limited to only the summative aspect of L2 writing assessment regarding the use of AWE systems. Future studies can be conducted to investigate the effectiveness of such systems in providing feedback to students' written productions and how these systems can be implemented in L2 contexts. In the era of technological advancements, the use of AWE-like systems can help to create more favorable learning environments and support teachers in designing more effective writing classes.

References

Alpar, R. (2010). *Uygulamalı istatistik ve geçerlik-güvenirlik: spor, sağlık ve eğitim bilimlerinden örneklerle*. Detay Yayıncılık.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. *The Journal of Technology, Learning and Assessment*, *4*(3), 1-31.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). Working inside the black box: Assessment for learning in the classroom. *Phi delta kappan*, *86*(1), 8-21.

Burstein, J., & Chodorow, M. (1999). Automated essay scoring for nonnative English speakers. Retrieved from https://origin-www.ets.org/Media/Research/pdf/erater_acl99rev.pdf

Cai, L. J. (2013). Students' perceptions of academic writing: A needs analysis of EAP in China. Language Education in Asia, 4 (1), 5-22.

Cotos, E. (2010). *Automated writing evaluation for non-native speaker English academic writing: The case of IADE and its formative feedback.* Unpublished Doctoral Dissertation. Iowa: Iowa State University, Faculty of Applied Linguistics and Technology.

Earl, L., & Katz, S. (2006). Rethinking classroom assessment with purpose in mind. *Winnipeg, Manitoba: Western Northern Canadian Protocol*.

El Ebyary, K., & Windeatt, S. (2010). The impact of computer-based feedback on students' written work. *International Journal of English Studies*, *10*(2), 121-142.

French, A. (2020). Academic writing as identity-work in higher education: forming a 'professional writing in higher education habitus'. *Studies in Higher Education*, 45(8), 1605-1617.

Gardner, J. (2012). *Assessment and learning*. London: Sage.

Lee, I. (2017). *Classroom writing assessment and feedback in L2 school contexts*. Singapore: Springer.

Lerner, N. D. (1996). *Teaching and learning in a university writing center*. Unpublished dissertation, Massachusetts: Boston University, Department of Humanities and Social Sciences.

Long, R. (2013). A review of ETS's Criterion online writing program for student compositions. *The Language Teacher*, 37(3), 11-18.

McNamara, T. (2000). *Language testing*. Oxford: Oxford University Press.

Pearson, P. D. (1998). Standards and assessments: Tools for crafting effective instruction? In J. Osborn & F. Lehr (Eds.), Literacy for all: Issues in teaching and learning (pp. 264–286). New York, NY: Guilford Press.

Powers, D. E., Burstein, J. C., Chodorow, M. S., Fowles, M. E., & Kukich, K. (2002). Comparing the validity of automated and human scoring of essays. *Journal of Educational Computing Research*, 26(4), 407-425.

Ramineni, C. (2013). Validating automated essay scoring for online writing placement. *Assessing Writing*, *18*(1), 40-61.

Richards, J. (1974) A non-contrastive approach to error analysis. In Richards, J. (ed.) *Error Analysis: Perspectives on Second Language Acquisition*. London: Longman.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, 25(4), 465-493.

Scriven, M. (1991). *Evaluation thesaurus*. Los Angeles, CA: Sage.

Shermis, M. D., Koch, C. M., Page, E. B., Keith, T. Z., & Harrington, S. (2002). Trait ratings for automated essay grading. *Educational and Psychological Measuremnt*, 62(1), 5-18.

Silva, T. (1993). Toward an understanding of the distinct nature of L2 writing: The ESL research and its implications. *TESOL quarterly*, 27(4), 657-677.

Wilson, J., & Roscoe, R. D. (2020). Automated writing evaluation and feedback: Multiple metrics of efficacy. *Journal of Educational Computing Research*, *58*(1), 87-125.