The Reading Matrix: An International Online Journal
Volume 21, Number 2, September 2021

Automated L2 Writing Performance Assessment: A Literature Review

Elif Sari
Karadeniz Technical University

Turgay Han
Ordu University

ABSTRACT

Providing both effective feedback applications and reliable assessment practices are two central issues in ESL/EFL writing instruction contexts. Giving individual feedback is very difficult in crowded classes as it requires a great amount of time and effort for instructors. Moreover, instructors likely employ inconsistent assessment procedures, which poses a threat to the reliability of the assessment results. Automated Writing Evaluation (AWE) systems were developed to address these issues and have been subjected to a number of research studies so far. This paper presents a literature review regarding the development of AWE systems and the studies that were conducted to investigate the use of these systems for teaching and assessing writing in the last two decades. Based on the review of previous research, it is suggested that more studies are carried out to investigate the effectiveness of automated feedback when it is integrated with teacher feedback and the reliability of automated scoring in classroom-based writing assessment contexts.

INTRODUCTION

Written language can give a more accurate reflection of one's linguistic competence by measuring aspects of language constructs that multiple-choice questions cannot. Therefore, writing is an effective tool to evaluate the proficiency of English for Speakers of Other Languages (ESOL) in both high- and low-stakes language tests (Horning, 1987; Williamson, Xi, & Breyer, 2012). In this sense, responding to students' writing (e.g., feedback) and assessing students' writing proficiency (e.g., scoring) are two central concerns in second or foreign language writing (Hyland, 2003). This paper reviews the literature in relation to the problems in providing individualized feedback on students' written products and assessing writing, and the development of AWE systems as a solution to these problems. It also reviews the research studies which investigated the feedback and scoring functions of these systems.

Feedback in Writing

Feedback is generally regarded as an integral part of ESOL writing instruction (Parr & Timperley, 2010). The role of feedback in writing gained popularity with the process approach that built on "Flower and Hayes's (1981) cognitive theory of writing, which emphasized the need for writers to produce multiple drafts, encouraging teachers to provide feedback and suggest revisions on drafts during the process of writing itself" (Hyland, Conesa, & Cerezo, 2016, p.433). According to this approach, effective writing requires an iterative process which contains writing, receiving informative feedback from a teacher or reader, revising based on this feedback, and then repeating this process again and again (Attali, 2004; Burstein, Chodorow, & Leacock, 2003).

In a typical ESL/EFL writing classroom, an instructor is mainly responsible for giving feedback to all students. The process of revising students' written products and providing corrective feedback on errors is not cost-effective for writing teachers, especially in crowded classes (Chen, Chiu, & Liao 2009; Dikli, 2006). Generally, teachers avoid assigning writing tasks, or they cannot provide individual feedback on each written product. When they give feedback, they only provide feedback for structural errors (e.g., grammar, spelling, and punctuation) as they do not have enough time to focus on content and organization (Yagelski, 1995). There are also disparities among instructors regarding the type of feedback they employ (e.g., direct/indirect, form-focused/content-focused). Their choice is generally based on their prior teaching experiences and personal views about the effectiveness of a specific error correction method rather than on any experimental evidence that shows one type of error correction is superior to the others. It is also important to consider that "the ability to provide fair and detailed feedback on writing demands a level of skill and training that is beyond the capacity of many instructors" (Warschauer & Ware, 2006, p.158).

Writing Assessment

Although direct assessment is regarded as an effective way to measure students' writing ability, it has two main disadvantages: 1) the evaluation of essays requires longer time than the evaluation of multiple-choice tests; and 2) it is difficult to provide consistency, due to a variety of factors, among or within raters when scoring writing samples (Huot, 1990). These factors include the writing task (e.g., the topic or prompt, discourse mode, rhetorical context, and input), the rater(s) (e.g., the rater's first language and professional background, rating and teaching experience, personality, and training), the rating scale (e.g., holistic and analytic), the scoring process, test-takers themselves, and the features of the written text (Hamp-Lyons, 1990; Huot, 2002; McNamara, 1996). Teachers, researchers, and decision-makers must be aware of these factors because they trigger the problems of reliability and validity in writing assessment (Huang, 2008; Hyland et al., 2016).

Reliability and validity are two concepts that most influence the quality of an assessment procedure (Hyland, 2003). Reliability is the consistency of test takers' scores when they are tested on different occasions, evaluated through different tasks, or scored by different raters (Johnson, Penny, & Gordon, 2009). Various factors can affect a test taker's performance, such as the environment in which the test is administered, the instructions provided to test takers, the writing topic, the genre, the time of the day, etc. A writing test is regarded as reliable on condition that it minimizes the effects of these factors. The other dimension of reliability deals with the consistency in rating students' writing because writing assessment is based on subjective judgements (Hamp-Lyons, 1990; McNamara, 1996). Different raters may assign different scores to the same essay (i.e., inter-rater reliability), or the same rater may assign different scores to the essays which are of the same quality (i.e., intra-rater reliability) (Homburg, 1984). Therefore, it is difficult to make reliable and fair inferences about the test-taker's writing performance on the basis of her/his test score, which will decrease the validity of scores. Validity refers to the accuracy of interpretations made based on the test scores (Bachman, 1990). In other words, validity refers to "the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores" (Messick, 1989, p. 39). Although getting consistent scores from a test does not ensure that the test measures what it asserts to measure, reliability is a prerequisite for validity (Popham, 1981). Therefore, scoring reliability should be regarded "as a cornerstone of sound performance assessment" (Huang, 2008, p. 202).

In order to increase the reliability and validity of scores, it is suggested that two or more raters are involved in the writing assessment procedure after they receive rater training to interpret a specific rating scale in a consistent way (Blood, 2011). However, this is difficult to apply in most situations since it is not time-efficient and cost-effective (Attali & Burstein, 2006). Additionally, raters may have some unconscious biases that are resistant to being corrected through training (Blood, 2011).

The Development of Automated Writing Evaluation (AWE) Systems

To overcome the reliability and fairness problems in ESOL writing assessment, the implementation of technology in ESOL performance assessment was explored by Ellis Page in the 1960s. Page noticed that reviewing dozens of writing papers posed an obstacle for teachers to assign more writing tasks to their students, and believed that using technology to provide immediate feedback on writing would help students improve their writing skills (Shermis, Burstein, Higgings, & Zechner, 2010). In addition, there was a need for a standardized and economical procedure for assessing students' writing proficiency because measuring writing skill is a time-consuming, labour-intensive, and costly process when it is conducted by human raters (Chung & O'Neil, 1997; Myers, 2003). Therefore, Page (1967) developed Project Essay Grade (PEG) with the support of the College Entrance Examination Board (CEEB). PEG used multiple regression analysis to relate measurable features of a text, such as average sentence length, average essay length, number of prepositions, and number of commas, to those in a corpus of essays on the same topic which had been scored by human raters (Page, 1967; Shermis, Mzumara, Olson, & Harrington, 2001).

Even though PEG yielded high scoring reliability, it was not regarded as a practical application because the technology was not developed and accessible enough at that time. Moreover, it was criticized as it focused on surface structures of writing, ignored content related features, and was vulnerable to cheating such as writing essays with more words and commas (Dikli, 2006). Since the introduction of PEG, the field of AWE has witnessed much development with the aim of increasing the reliability and validity of AWE systems as writing evaluation tools (Attali, 2004; Dikli, 2006). With the advent of microcomputers and the Internet in the early 1980s, the use of technology in writing evaluation came to the fore and a second product, called the Writer's Workbench was launched (Attali, 2004; Shermis et al., 2010). The Writer's Workbench brought a new dimension in the field by providing feedback regarding the writing quality instead of scoring essays despite its limited capacity to identify quality (Warschauer & Ware, 2006). In the 1990s, three main essay scoring engines which are popular at present were released. One of these systems is e-rater which was developed by the Educational Testing Service (ETS) and the other is Intellimetric developed by Vantage Learning. Both of these scoring systems use Artificial Intelligence and Natural Language Processing methods to extract some features of writing (e.g. syntactic variety or the organization of ideas) from the training essays which were pre-scored by expert human raters. Then, they employ regression analysis to determine the best combinations of these features in order to imitate the scores assigned by expert human raters. Finally, these combinations are coded into the computer program to assess new essays (Attali & Burstein, 2006; Chodorow & Burstein, 2004). The third scoring system is Intelligent Essay Assessor, which was developed by academics and supported by Pearson Knowledge Technologies. Unlike e-rater and Intellimetric, it uses latent semantic analysis, which means comparing the semantic meaning of a target essay with a corpus of textual information on a similar topic (Landauer, Laham, & Foltz, 2003).

The developments in technology and the increasing importance of providing feedback on students' writing inspired the scoring systems to develop some programs which are appropriate for providing feedback in classroom setting; for example, Criterion by ETS, MyAccess by Vantage Learning, and WriteToLearn by Pearson Knowledge Technologies (Chen & Cheng, 2008; Chodorow & Burstein, 2004; Dikli, 2006; Warschauer & Grimes, 2008). Table 1 shows the AWE systems and their instructional applications that are widely used by testing companies, universities and public schools (Warschauer & Ware, 2006, p.3).

Table 1. Profiles of AWE Systems and Their Instructional Applications

Company	**Software Engine**	**Evaluation Mechanism**	**Commercial Product**	**Scoring**	**Feedback**
Vantage Learning	Intellimetric	Artificial Intelligence	MY Access!	Holistic and component scoring	Limited individualized feedback
Educational Testing Service	E-rater and Critique	Natural Language Processing	Criterion	Single holistic score	Wide range of individualized feedback
Pearson Knowledge Technologies	Intelligent Essay Assessor	Latent Semantic Analysis	WriteToLearn	Holistic and component scoring	Limited individualized feedback

AWE systems use computer technology to take a written text through a web page and provide feedback for writers' errors in various dimensions (e.g., grammar, vocabulary, mechanics, organization, etc.) and a score reflecting the overall quality of the written work within seconds (Chung & O'Neil, 1997; Hamp-Lyons, 2001; Shermis & Burstein, 2003). As well as providing formative feedback and scoring, these programs offer a variety of supporting features such as model essays, graphic organizers, rubrics, dictionaries, and thesauruses (Warschauer & Ware, 2006; Warschauer & Grimes, 2008). Although these programs were developed originally for English native speakers, they have added some features (e.g., multilingual feedback systems) that make them suitable for ESOL students (Warschauer & Ware, 2006).

AWE systems have both merits and drawbacks. On the one hand, AWE systems can provide a fast and cost-effective scoring with less effort to recruit trained human raters, and therefore making the direct assessment of writing skill as advantageous as the assessment of writing through multiple-choice tests (Williamson et al., 2012). For example; when TOEFL iBT was first released, it included two essays that were scored by two human raters. It was challenging to recruit, train, and maintain a pool of qualified raters and to report scores in a timely manner. After the development of e-rater, it was used instead of one of the human raters, which alleviated the challenges of the previous scoring procedure (Attali & Burstein, 2006; Enright & Quinlan, 2010). Students can also gain more responsibility and autonomy in their writing process since they can submit multiple drafts of a writing task and receive feedback anywhere and anytime, even when there is no access to human support (Wang & Goodman, 2012 cited in Liao, 2015). AWE systems allow students to practice independently, so students need less input from their teachers (Steinhart, 2001). Additionally, receiving instantaneous feedback in writing instruction motivates students to spend more time doing revisions in their writing tasks while their statements are still fresh in their minds (Spencer & Louw, 2008). Moreover, teachers can spend more time providing feedback related to the content and organization of writing. They can also communicate with their students on producing ideas because they are freed from marking papers for surface-level errors, such as grammar or mechanics (Burstein et al., 2003). Finally, the use of AWE can impact the improvement of students' writing quality because teachers can assign more writing tasks to their students and they can have more effective interaction with their students (Spencer & Louw, 2008; Warschauer & Ware, 2006).

On the other hand, even though all of these AWE systems are trained using essays scored by human raters to make predictions on a target essay, they cannot evaluate a text in the same way as human readers do. For example, e-rater evaluates the organization and development of an essay based on the existence or absence of the introduction, thesis statements, supporting details, and conclusion. However, human readers consider whether the thoughts are presented in a logical order, the introduction covers the text, and the conclusion adequately encapsulates the message that the writer intends to convey. In addition, human readers bring their world knowledge and inferencing skills into the evaluation process, so they can understand allusions, humour, or irony that cannot be perceived by e-rater (Weigle, 2013). Furthermore, the use of AWE systems in the classroom may lead students to put emphasis on mechanical features and grammatical correctness by ignoring that the purpose of writing is to convey meaning since AWE systems are more likely to provide feedback based on these traits (Cheville, 2004). If learners know that they are writing for asocial machines, they can make an effort to use the right words in the right chains without regarding writing as a means of social interaction, which might direct them to use formulaic expressions (Ericsson & Haswell, 2006).

The use of AWE in ESL/EFL writing contexts has been subject to a great deal of studies for the last two decades because many teachers and administrators from K-12 classrooms, colleges, and universities began to use AWE systems as educational tools (Stevenson & Phakiti, 2014). These studies focused either on the scoring or the feedback function of AWE systems. These studies are reviewed in the next section.

RESEARCH INTO AUTOMATED WRITING EVALUATION

This section reviews 44 studies that have investigated the feedback and scoring functions of different AWE systems. Table 2 illustrates a summary of these studies.

Table 2. Studies Reviewed

	Focus of the Study	Number of Studies Reviewed	Studies Reviewed
Automated Feedback	The Impact of Automated Feedback on Revision and Improvement in Writing	13	Attali, 2004 Choi, 2010 Chou, Moslehpour, & Yang, 2016 Dikli, 2014 Ebyary & Windeat, 2010 Kellogg, Whiteford, & Quinlan, 2010 Lai, 2010 Liao, 2015 Rock, 2007 Tang & Rich, 2017 Wang, 2015 Wang, Shang, & Briody, 2013 Wilson & Czik, 2016
	Students' and/or Teachers' Perceptions towards AWE	12	Chen & Cheng, 2008 Dikli, 2006, 2014

			Fang, 2010 Lai, 2010 Li, Link, & Hegelheimer, 2015 Link, Dursun, Karakaya, & Hegelheimer, 2014 Maeng, 2010 Wang, 2015 Warschauer & Grimes, 2008 Grimes & Warschauer, 2010 Tsuda, 2014
Automated Scoring	The Reliability and Validity of Automated Scoring in Large-scale Standardized Tests	13	Attali, 2007 Attali & Burstein, 2006 Burstein & Chodorow, 1999 Burstein et al., 1998 Chodorow & Burstein, 2004 Elliot, 2001 Foltz, Kintsch, & Landauer, 1999 Landauer, Foltz, & Laham, 1997 Petersen, 1997 Powers et al., 2002a Powers et al., 2002b Shermis et al., 2002 Wang & Brown, 2007
	The Reliability of Automated Scoring in Classroom-based Writing Tests	6	Bridgeman, Trapani, & Attali, 2009 Ebyary & Windeat, 2010 Hoang & Kunnan, 2016 James, 2006 Li et al., 2015 Liu & Kunnan, 2016
Total		44	

Research into Automated Feedback

The studies investigating the automated feedback provided by the AWE systems for instructional purposes have focused on two different aspects of it: a) the impact of automated feedback on students' revision processes and writing improvement (Attali, 2004; Choi, 2010; Chou et al., 2016; Dikli, 2014; Ebyary & Windeat, 2010; Kellogg et al., 2010; Lai, 2010; Liao, 2015; Rock, 2007; Tang & Rich, 2017; Wang, 2015; Wang et al., 2013;Wilson & Czik, 2016); and b) students' and/or teachers' perceptions regarding the effectiveness of automated feedback (Chen & Cheng, 2008; Dikli, 2006, 2014; Fang, 2010; Lai, 2010; Li et al., 2015; Link et al., 2014; Maeng, 2010; Wang, 2015; Warschauer & Grimes, 2008; Grimes & Warschauer, 2010; Tsuda, 2014).

Studies Investigating the Impact of Automated Feedback on Students' Revision Processes and Writing Improvement

Several studies have investigated how automated feedback affects students' revision processes and writing improvement (Attali, 2004; Choi, 2010; Chou et al., 2016; Dikli, 2014; Ebyary & Windeat, 2010; Kellogg et al., 2010; Lai, 2010; Liao, 2015; Rock, 2007; Tang & Rich, 2017; Wang, 2015; Wang et al., 2013; Wilson & Czik, 2016). Some of these studies have examined the impact of automated feedback on students' writing scores and revealed contradicting results (Choi, 2010; Chou et al., 2016; Ebyary & Windeat, 2010; Rock, 2007; Tang & Rich, 2017; Wilson & Czik, 2016). For example, Rock (2007) investigated to what extent the short-term use of Criterion affected students' writing scores. The students were randomly assigned to the treatment group who used Criterion as a supplemental instructional tool and to the control group who continued their traditional writing instruction. The results showed that the students in the treatment group received higher analytic scores than the students in the control group on their essays which they wrote at the end of the treatment. Although this difference was not so large, it was statistically significant. On the other hand, no difference was found between the two groups regarding their holistic scores. This result might be attributed to the limited time allowed for the study. In addition, the analytic scores demonstrated that Criterion improved the mechanical aspects of student essays more than the other aspects. In a longer term study, Ebyary and Windeat (2010) found a significant improvement in trainee EFL teachers' holistic scores between the two drafts (first draft and final draft) and among the four submissions in total. In addition, regular and timely feedback provided by the program increased the participants' motivation to write more. Similarly, Chou et al. (2016) investigated the impact of using My Access on pre-intermediate EFL college students' writing quality and self-correction of errors by comparing their submissions. The results revealed improvement between submissions in terms of holistic and analytic scores and text length. Using My Access motivated the students to revise more and write longer than the pen-and-paper method. Additionally, the students expressed mostly positive views towards the use of technology in EFL writing instruction. In their quasi-experimental mixed-methods study, Tang and Rich (2017) investigated the effects of an AWE system, Writing Roadmap, on students' writing processes and writing test scores in secondary schools and universities in China. Both secondary school students and university students participated in the study. The results of the pre- and post-tests revealed that for the college-level sample, the experimental group students who received automated feedback achieved greater writing improvement than the control group students who received teacher feedback. For the high school sample, the difference between the two groups of students in terms of writing proficiency was smaller, but the students in the experimental group received higher scores on their post-test than the students in the control group.

Contrary to the aforementioned studies, Wilson and Czik (2016) found no significant difference between the two groups of students who were assigned either to a teacher+automated feedback condition or to a teacher feedback-only condition. They also found that the amount of feedback the teachers provided in both conditions was more or less the same, but the feedback given to the students in the teacher+automated feedback condition was mostly related to higher level writing skills (e.g., ideas and elaboration, organization, and style). The teachers stated that they spent less time providing feedback in the teacher+automated feedback condition than they spent in the teacher feedback-only condition. In addition, the students who received both automated and teacher feedback were more motivated and persistent in their writing. Choi (2010) investigated the impact of AWE on students' writing quality and accuracy in his dissertation study through three different AWE integration models: No-AWE (only teacher feedback), Optional-AWE (students use AWE at their discretion), and Integrated-AWE (automated feedback+teacher feedback). It was concluded that although the Integrated-AWE group achieved the most improvement from first to revised draft of a writing assignment, no significance difference was found across the three groups in terms of improving their holistic writing scores from pre- to post-test. However, the ESL group achieved more progress than the EFL group regarding their holistic writing quality, which indicated that English learning context was more effective on the students' writing improvement than the integration model.

Some other studies have examined the impact of automated feedback on students' writing errors and showed that automated feedback helped students reduce their errors (Attali, 2004; Kellogg et al., 2010; Liao, 2015; Wang et al., 2013). Attali (2004) examined whether students used the feedback provided by Criterion to correct their errors by looking at the improvement in their scores from the first to the last submissions. Results showed that the feedback provided by Criterion was effective for students to improve their essays by reducing their errors of grammar, usage, mechanics, and style and increasing the length of their essays. Additionally, the rate of occurrence of background and conclusion elements, and the number of main points and supporting ideas elements increased from the students' first to last submissions. In their study, Kellogg et al. (2010) examined to what extent the automated feedback and scoring provided by the Criterion program influenced students' writing performance in a freshman composition course. The students' holistic scores showed no significant gains of feedback in general, but the students who received continuous automated feedback reduced their number of errors in the categories of mechanics, usage, and grammar. On the other hand, intermittent automated feedback was found to be ineffective in reducing the number of errors on the transfer test. Further, Wang et al., (2013) examined the impact of AWE (CorrectEnglish) on Taiwanese EFL freshmen's writing performance with regards to accuracy. The results of the essay analysis showed that the students in the experimental group made fewer errors on their written products after the AWE treatment. Two years later, Liao (2015) found that the use of Criterion helped students reduce their number of errors in four identified types of grammatical errors (e.g., fragments, run-on sentences, subject-verb disagreement, and ill-formed verbs) in both revisions and new texts although the effect was changeable for each error category.

Studies Investigating Students' and/or Teachers' Perceptions of Using AWE

Several studies have examined students' and/or teachers' perceptions regarding the use of automated feedback in writing instruction and revealed mixed results (Chen & Cheng, 2008; Dikli, 2006, 2014; Fang, 2010; Lai, 2010; Li et al., 2015; Link et al., 2014; Maeng, 2010; Wang, 2015; Warschauer & Grimes, 2008; Grimes & Warschauer, 2010; Tsuda, 2014). Grimes and Warschauer (2010) concluded that using My Access in writing classes facilitated classroom management since the immediate feedback given by the program increased the students' motivation to write and revise. The students felt themselves more confident in writing as they found machine judgement less threatening than human judgement. In addition, the use of AWE allowed teachers to spend time more on higher-level concerns than mechanics. The researchers concluded that the use of AWE may not improve ineffective teaching, but it can make effective teaching much more effective. Later, Link et al. (2014) and Li et al. (2015) obtained similar results to those found by Grimes and Warschauer (2010). The two studies demonstrated that Criterion allowed instructors to provide more detailed feedback in terms of content and organization as they left correcting grammatical and mechanical errors to the students' responsibility. The instructors thought the program was effective and satisfactory as it increased student motivation and decreased instructor workload. However, Li et al. (2015) reported that the instructors thought the feedback on content was unhelpful and the scoring ability of the program was problematic. In addition, Link et al. (2014) stated that the effectiveness of the online program was determined by the instructors' willingness to implement it, the way in which the instructors implemented it, and the ability of the instructors and the students to overcome the technological difficulties.

In the same vein, Tang and Rich (2017) found that students were motivated to write and revise more when they used the computer program as they actively participated in the assessment process and took responsibility for their own progress. More importantly, students learned to correct their own errors as a result of their constant interaction with the automated feedback. In addition, teachers saved time to provide feedback on content and organization of the written text rather than language form. This study also indicated that the way of introducing technology, teacher training and guidance, and student training enhance the efficacy of AWE.

Fang's study (2010) indicated that EFL students were positive about using My Access as a writing instruction tool as it increased their revision of their assignments. However, their attitude toward using My Access as a writing scoring tool was less positive. In addition, the students asserted that the use of this AWE tool increased their writing skill in terms of form rather than content. In the same year, Maeng (2010) investigated the perceptions of secondary English teachers regarding the effectiveness of using Criterion in a teacher training program. The teachers viewed Criterion as a helpful tool to develop their writing skills even though they thought that the quality of the automated feedback was inadequate. Similarly, in Tsuda' study (2014), most of the students regarded Criterion as an effective instruction tool to improve their writing skills as it allowed them to get rid of their repeated mistakes. However, they expressed some criticisms regarding the technical problems of the program, the inefficiency of the program in providing feedback in discourse and content level, and the limitation of the program in supplying alternative ways or advice about how to correct the errors.

Contrary to the aforementioned studies which demonstrated positive perceptions, some studies revealed neutral or negative perceptions towards the use of automated feedback and scoring. For example, Dikli (2006) examined university level students' perceptions of using MY Access in their writing classes. The results showed that while the teacher feedback was shorter and more specific, the automated feedback was long, general, and redundant. The students in the teacher feedback group received less feedback from their teacher but they used most of the feedback in their drafts. However, the students in the AWE feedback group sometimes did not read the feedback provided by the program because they found the feedback unnecessary and confusing. The students spent much time dealing with the mechanical problems in their essays. Therefore, they could not concentrate on other traits of their writing. Both the students and their teacher stated that human interaction was absent in the automated feedback condition and the students favoured teacher feedback rather than the automated feedback. In another study, Dikli (2014) compared Criterion feedback to teacher feedback and concluded that teachers gave more useful and accurate feedback than the online program. In addition, Criterion missed some errors that were identified by the teacher. Although some students noticed the weaknesses of the automated feedback, they generally trusted the program and held positive views regarding its use in writing instruction, especially when it is used in conjunction with teacher feedback.

In their naturalistic classroom-based study, Chen and Cheng (2008) implemented My Access in three EFL college writing classes. The instructors integrated the program in their classes in three different ways. The results showed that the students in the three classes did not favour AWE due to some limitations in the design of the program. It was found that the program was more effective when it was used to help students in their early drafting and revising process, followed by teacher and peer feedback in the later stages. On the other hand, using AWE as a main writing coach with minimal teacher feedback frustrated the students and limited their improvement in writing. They suggested that human feedback and assessment is necessary in AWE learning environments.

Next, Warschauer and Grimes (2008) concluded that nearly half of the participant students edited their papers more when they used the AWE program, but most of the editing was at the word or sentence level. Also, they found that there were differences among the students in using the program because of their socioeconomic status. Students with high-socioeconomic status were able to use the program better thanks to their keyboard skills, computer and Internet access at home, and language and literacy background. As a result, the researchers suggested that the effectiveness of AWE programs depends on teachers' ability to integrate them into their teaching programs in a way that fits their students' needs, socioeconomic status, and computer experience.

Two years later, Lai (2010) compared peer feedback and automated feedback in terms of how students used these two types of feedback in their revisions and what perceptions the students held toward the two types of feedback. The participants were 22 college-level EFL learners (English majors) in Taiwan. The AWE system used in this study was My Access. The results of the questionnaire and interviews demonstrated that the college-level EFL learners favoured peer feedback over automated feedback with regards to process, product, and perspective. This finding supported the positive impact of social interaction with peers. The students accepted their peers as real audiences and valued their comments on their writings. In addition, they found automated feedback too general, vague, and fixed while they thought peer feedback was more direct and explicit. Another reason for preferring peer feedback was that the students felt less comfortable while writing their tasks on the computer and they had problems with the Internet connection.

Furthermore, in Wang's study (2015) most of the students believed that their writing skills improved after using Criterion but they stated that it was faulty to attribute their improvement only to Criterion. They found their teacher's guidance vital during the treatment, especially when they had difficulty in understanding the feedback provided by Criterion or they found the diagnostic message misleading or incorrect. Regarding the effectiveness of the diagnostic feedback provided by Criterion, students found the feedback on grammar and usage more beneficial than the feedback on mechanics and style. With reference to the correctness of the diagnostic feedback messages, students reported that some messages were incorrect and confusing, and some basic errors in grammar and usage were missed by the program.

Research into Automated Scoring

Several studies have been carried out to investigate the reliability and validity of the scores given by AWE systems in large-scale standardized tests like TOEFL, GRE, or TWE (Attali, 2007; Attali & Burstein, 2006; Burstein & Chodorow, 1999; Burstein et al., 1998; Chodorow & Burstein, 2004; Elliot, 2001; Foltz et al., 1999; Landauer et al., 1997; Petersen, 1997; Powers et al., 2002a; Powers et al., 2002b; Shermis et al., 2002; Wang & Brown, 2007). On the other hand, some other studies examined the reliability of automated scoring in classroom-based writing assessment contexts (Bridgeman et al., 2009; Ebyary & Windeat, 2010; Hoang & Kunnan, 2016; James, 2006; Li et al., 2015; Liu & Kunnan, 2016).

Studies Investigating the Reliability and Validity of AWE Systems in Large-scale Standardized Tests

Most of the studies investigating the reliability and validity of the scores provided by AWE systems have centred on the agreement rates between human raters and automated raters (Burstein et al., 1998; Elliot, 2001; Foltz et al., 1999; Landauer et al., 1997; Shermis et al., 2002). The correlations between the scores assigned by an AWE program and a human rater were compared to the correlations between scores assigned by two human raters. The studies that focused on the agreement of automated scores with human scores stemmed from the assumption that human-assigned scores are valid enough to accept as the "gold standard" which enables us to make inferences about the validity of automated scores (Powers et al., 2000). These studies have shown that the correlations between human raters and e-rater are approximately as high as the correlations between two human raters (Attali & Burstein, 2006).

The agreement between human scores and automated scores is valuable as human raters can evaluate the content and logical structure of a written product that automated scoring systems may fail to evaluate, but it cannot be the only indicator of the reliability and validity of these systems. Agreement results provide inadequate information about the construct validity of AWE systems. One way of investigating the validity of AWE systems is to compare automated scores with other scores from different tests measuring the same or similar construct (alternate-test); for example, the scores of a multiple-choice test of writing (Attali, 2007; Bridgeman et al., 2012). In 1997, Petersen implemented this procedure to examine the validity of the PEG system using the essays of prospective teachers who were attending a teacher certification program, the Praxis Series: Professional Assessment for Beginning Teachers. She measured the correlation of e-rater and human rater scores with the scores of the multiple-choice writing subtest. She found that the correlation of e-rater scores with the writing subtest scores was .47; however, the correlation of human rater scores with the same scores was .45. Additionally, she analysed the correlation of automated and human rater scores assigned on essays with the scores of other subtests in the same general ability test, such as reading and mathematics. She concluded that the correlations of e-rater scores with the scores of reading and mathematics subtests were .39 and .30, respectively. However, the correlations of human rater scores with the same subscores were slightly higher, .43 and .36, respectively.

Attali and Burstein (2006) stated that the analysis of human-machine agreement on a single essay reveals some problems as reliability entails the consistency of scores drawn from various administrations. Thus, they designed various parallel prompts for 6th to 12th grade students, so the essays written in response to each prompt could be used as alternative forms. Their dataset consisted of 4000 essays from 2000 students on two alternative topics. According to the results, e-rater scores had higher alternate-form reliability than single human scores in six out of seven grades. Moreover, the overall alternate-form reliability of e-rater scores was almost the same as the average of scores assigned by human raters (.59 and .58, respectively). They also found a high true score correlation between e-rater and human rater scores (.97).

In the same vein, employing a multitrait-multimethod approach, Attali (2007) analysed the essays of 5,006 examinees from 31 countries who repeated the test twice through three types of analyses. First, correlational analysis was carried out between all scores obtained from the two tests, which allowed to test alternate-form reliability. Second, the correlations and reliabilities of the essay scores assigned by human rater and e-rater were analysed together with TOEFL subscores (structured writing, reading, and listening). Third, essay score correlations were compared to essay length. As in Attali and Burstein (2006), this analysis showed that the e-rater reliability (.71) was higher than the single human rater reliability (.54) and the double human rater reliability (.63). The reliabilities of TOEFL subscores (structured writing, reading, and listening) were around .80. In addition, the correlations between essay length and the average score of human raters and e-rater were 0.57 and 0.61, respectively.

Powers et al. (2002a) investigated the relationship of both automated and human rater scores on essays written in the Graduate Record Examinations (GRE) to some other indicators of writing skill that are beyond the context of the test to be mentioned, such as academic, outside, and self-reported success regarding individual writing skill. Their database consisted of 101 to 149 essays written as a response to 20 argument and issue prompts. The results revealed that the correlation between automated scores and non-test indicators of writing skill was similar to that of the scores assigned by human raters although that kind of correlation was partly higher for human rater scores than automated scores.

Another study by Powers et al. (2002b) investigated the sensitivity of AWE systems to the extraneous features of test-takers' writing skill which pose a threat to the construct validity of automated scoring. They invited 27 writing experts to trick e-rater into assigning scores higher or lower than they deserved. The experts were required to write two complementary essays in response to each of the two GRE writing prompts. Their texts were scored both by e-rater and two trained human raters. The difference between e-rater and human scores elicited to what extent e-rater could be deceived. E-rater was found to be vulnerable to the experts' tricks to obtain scores higher than they deserved. This detection suggests that automated scoring systems may be mistaken so they should be used together with human scoring in high-stakes assessment contexts.

Since automated scoring systems were originally designed to evaluate writing produced by native speakers of English, it was crucial to investigate whether these systems function properly for non-native speakers of English. Burstein and Chodorow (1999) investigated the human rater and e-rater agreement for non-native speakers of English on roughly 1100 essays written as a response to the two prompts in the Test of Written English (TWE). The results revealed that e-rater had the ability to evaluate the syntactic and discourse structure of non-native speakers' written products as well. Although native and non-native speakers obtained considerably different scores from both e-rater and human rater, the agreement between the two kinds of scoring was as high as it was obtained for native speakers. The study also showed that the language group significantly affected human or machine scoring. Arabic and Spanish speakers got higher scores from human raters than from e-rater whereas essays of Chinese speakers got higher scores from e-rater. The differences were slight, but the most significant difference was for Chinese speakers with a standard deviation of 0.48.

In another study, Chodorow and Burstein (2004) examined the effect of essay length on the automated scores that were assigned to TOEFL essays written by examinees that were of different ethnicities. Using a mixed model repeated measures ANOVA, they analyzed 265 training essays for each of seven prompts. Results showed that e-rater and human scores differed across language groups for only one essay prompt on which Arabic and Japanese speakers received higher scores from e-rater, but Spanish speakers received the same scores both from human rater and e-rater. They also found that essay length was a factor that could explain the variance in human rater and e-rater scores. However, the later version of e-rater was less affected by essay length than its early version. In addition, e-rater and human raters yielded the same forms of differences across native language groups when essay length was controlled.

Furthermore, Wang and Brown (2007) examined the validity and usefulness of AWE in scoring different dimensions of essays in large-scale placement tests through a correlational study. They correlated the holistic scores assigned by human raters and Intellimetric on the essays written by 107 Hispanic examinees for the WritePlace Plus test and found no significant correlation. They concluded that human scores and machine scores were consistent only in sentence structure which opposed the findings of Vantage Learning (2000) that claimed high consistency for focus, content, organization, and style.

Studies Investigating the Reliability of AWE Systems in Classroom-based Writing Assessment

The studies investigating the use of automated scoring in classroom-based writing assessment (Bridgeman et al., 2009; Ebyary & Windeat, 2010; Hoang & Kunnan, 2016; Huang, 2014; James, 2006; Li et al., 2014; Liu & Kunnan, 2016) have frequently indicated lower agreement between automated scores and human rater scores than the studies conducted on the samples selected from high-stakes tests. For example, Bridgeman et al. (2009) conducted a study on the essays written by the eleventh grade students in an end-of-course test. The essays were holistically evaluated by two human raters and e-rater. The researchers found .84 agreement for human-human and .76 agreement for human-machine scoring. In the same vein, James (2006) examined the accuracy of automated scoring in a placement test. Writing samples were obtained from 60 students from the University Preparation Department of a post-secondary program. The essays were scored by Intellimetric and 11 native English instructors from the same institution. The faculty scorers had no training in assessing writing. The results revealed positive correlations between the scores assigned by human raters (between .45 and

.80). However, lower and narrower correlations were found between human rater scores and automated scores (between .40 and .61).

Next, Ebyary and Windeat (2010) obtained the writing samples written on four different topics by 24 trainee EFL instructors. The samples were scored by Criterion and two English language instructors using Criterion scoring scale with the aim of examining the accuracy of scores produced by Criterion. Pearson correlation was conducted to measure inter-rater reliability. While a significant inter-rater reliability was found between the automated scores and the scores provided by the first human rater (r = .83), a moderate reliability was found between the second rater and Criterion (r = .53). Four years later, Huang (2014) investigated the difference between human scoring and automated scoring. 103 essays written by 26 Taiwanese English majors on four different writing prompts were scored by Criterion and two human raters. The t-test analysis showed that there was a difference between the scores given by each human rater and Criterion. The average scores assigned by the program were higher than the average scores given by the human raters. The correlation test revealed that Criterion moderately correlated with the first human rater while it significantly correlated with the second human rater.

Li et al. (2014) investigated the use of the automated scoring in classroom-based formative writing assessment by analysing the correlation between the holistic scores assigned by Criterion and the holistic and analytic scores given by three ESL writing instructors in three college-level ESL writing courses. Correlation analysis revealed that there was a moderate or low consistency between the instructors' holistic scores and the scores assigned by Criterion depending on the task. Instructors expressed their neutral thoughts regarding their trust in the Criterion scores. Similarly, students held moderate trust towards the automated scores they received, but regarded the program as an effective motivator in the writing process. Therefore, instructors used Criterion scores as an indicator of students' writing quality in the pre-submission process and avoided using these scores as the only summative assessment tool in classroom-based writing instruction. In addition to the consistency of the holistic scores assigned by instructors and Criterion, the study examined the correlation between the instructors' analytic scores and the automated holistic scores. Organization and Correctness were found to be the two subcategories which revealed the highest correlation with automated scores among other subcategories, such as Material, Expression, and Paper Process.

More recently, Hoang and Kunnan (2016) examined the agreement between the scores assigned by My Access and human raters on the essays written by EFL and ESL learners as a response to three writing prompts. The correlational analyses showed a better agreement between the two human raters (.78) than the agreement between the human raters and My Access (.68). The means of the averaged scores revealed that My Access assigned higher scores (M = 4.09) to students' essays than the human raters (M = 3.76). Contrary to this study, Liu and Kunnan (2016) found that the scoring of WriteToLearn was more consistent but more severe than the human rater scoring when they compared the scoring performance of WriteToLearn against those of four trained human raters.

CONCLUSION

AWE was developed to reduce teachers' workload in writing instruction and to increase reliability and fairness in writing assessment. Even though AWE systems have been subject to a great number of studies for the last two decades, there are still some gaps in the AWE research. First, the studies investigating the effects of AWE systems on learners' writing development and the users' perceptions regarding the effectiveness of these systems demonstrated contradictory results since they employed different designs. Several studies employed within group designs in which a single group of students' performance was compared across submissions (e.g. Attali, 2004; Chou et al., 2016; Lai, 2010; Liao, 2015). In these studies, the lack of control group makes it uncertain whether the students' improvement is the result of the automated feedback they received because other factors (e.g., classroom instruction or developmental factors) could be the source of their writing improvement (Stevenson & Phakiti, 2014). Several studies used between group designs and compared automated feedback with a no-feedback condition which is a weak counterfactual against automated feedback (e.g. Cheng, 2017; Kellogg et al., 2010). Some others compared automated feedback with a teacher feedback condition (e.g. Wang et at., 2013; Rock, 2007), but these studies investigated the impact of using automated feedback in isolation of teacher feedback, which conflicts with the fact that AWE systems were developed to complement teacher feedback not to replace it (Burstein et al., 2003; Warschauer & Ware, 2006). In this sense, few studies have investigated the impact of AWE systems when they are used as instructional tools that supplement teacher feedback including a control group (e.g., Choi, 2010; Tang & Rich, 2017; Wilson & Czik, 2016). However, these studies do not elaborate the nature of the teacher feedback in the control groups and how automated feedback was integrated with teacher feedback in the experimental groups, which makes it difficult to know whether the feedback conditions were comparable between the groups. Furthermore, it is necessary to note that the effectiveness of AWE integration depends on several factors such as students' familiarity with technology, teachers' readiness and willingness to integrate technology in their instruction, and training both students and teachers on how to use the selected AWE program (Grimes & Warschauer, 2010; Warschauer & Grimes, 2008). However, the previous studies do not explain whether the teachers were informed enough about integrating technology into their teaching or the students had the necessary technological skills to use the AWE system they were required to use.

Second, most of the studies which showed that automated writing scoring is as reliable as human scoring were conducted on samples selected from high-stakes standardized tests (e.g., TOEFL, GRE, or TWE). The fact that whether automated raters provide a reliable and accurate writing assessment for low-stakes writing tasks and classroom-based writing tests is under-researched and a limited number of studies on this issue revealed conflicting results (Bridgeman et al., 2009; Ebyary & Windeat, 2010; Hoang & Kunnan, 2016; James, 2006; Li et al., 2015; Liu & Kunnan, 2016).

As a result, it is concluded that it is necessary to conduct rigorous experimental studies in order to attain more reliable results regarding the effectiveness of AWE systems as supplementary instruction tools. Future studies should consider the factors that contribute to the effective use of AWE systems (e.g., teachers' ability and willingness to integrate technology into their teaching, students' familiarity with using technology, and training on using AWE tools for both teachers and students) in their research designs. Future studies should also elaborate how they integrate AWE with teacher feedback in order to provide a concrete model of AWE-integrated writing instruction for teachers who are thinking of integrating technology into their writing classes. Additionally, more studies should be conducted to investigate the accuracy of automated scores in classroom-based regular writing assessment (e.g., assessing EFL students' writing tasks during the writing skill course) and classroom-based high-stake EFL writing assessment contexts (e.g., English proficiency exams for entering and exit ELT departments or writing tests applied for selecting students for international exchange programs) with the purpose of making contributions to the reliability of writing assessment practices.

ACKNOWLEDGEMENT

This study was derived from a PhD dissertation by Elif Sari under the supervision of Turgay Han and it was approved by the Faculty of Educational Sciences of Atatürk University in Turkey.

We would like to express our gratitude to TUBITAK (The Scientific and Technological Research Council of Turkey) for financial support provided through 2211-National Graduate Scholarship Programme.

REFERENCES

Attali, Y. (2004). *Exploring the feedback and revision features of Criterion*. Paper presented at the National Council on Measurement in Education (NCME). San Diego, CA.

Attali, Y. (2007). *Construct validity of e-rater in scoring TOEFL essays* (Research Report RR-07-21). Princeton, NJ: Educational Testing Service.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater V. 2. *The Journal of Technology, Learning and Assessment, 4*(3), 3-30.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Blood, I. (2011). Automated essay scoring: A literature review. *Studies in Applied Linguistics and TESOL, 11*(2), 40-64.

Bridgeman, B., Trapani, C., & Attali, Y. (2009, April). *Considering fairness and validity in evaluating automated scoring*. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME). San Diego, CA. Retrieved from http://www.ets.org/Media/Conferences_and_Events/AERA_2009_pdfs/AERA_NCME_2009_Bridgeman.pdf

Burstein, J., Braden-Harder, L., Chodorow, M., Hua, S., Kaplan, B., Kukich, K., ... & Wolff, S. (1998). Computer analysis of essay content for automated score prediction: A prototype automated scoring system for GMAT analytical writing assessment essays. *ETS Research Report Series, 1998*(1), i-67.

Burstein, J., & Chodorow, M. (1999). Automated essay scoring for nonnative English speakers. *Proceedings from the ACL99 Workshop on Computer-Mediated Language Assessment and Evaluation of Natural Language Processing*. College Park, MD.

Burstein, J., Chodorow, M., & Leacock, C. (2003, August). Criterion online essay evaluation: An application for automated evaluation of student essays. *Proceedings from the 15th Annual Conference on Innovative Applications of Artificial Intelligence*. Acapulco, Mexico.

Chen, C., & Cheng, W. (2008). Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes. *Language Learning & Technology, 12*(2), 94-112.

Chen, H. H. J., Chiu, S. T. L., & Liao, P. (2009). Analyzing the grammar feedback of two automated writing evaluation systems: My Access and Criterion. *English Teaching and Learning, 33*(2), 1-43.

Cheville, J. (2004). Automated scoring technologies and the rising influence of error. *English Journal, 93*(4), 47-52.

Chodorow, M., & Burstein, J. (2004). *Beyond essay length: Evaluating e-rater's performance on TOEFL essays* (Research report No. 73). Princeton, NJ: Educational Testing Service.

Choi, J. (2010). *The impact of automated essay scoring (AES) for improving English language learner's essay writing* (Unpublished PhD thesis). Charlottesville, VA: University of Virginia.

Chou, H. N. C., Moslehpour, M., & Yang, C. Y. (2016). My access and writing error corrections of EFL college pre-intermediate students. *International Journal of Education, 8*(1), 144-161.

Chung, G., & O'Neill, H. (1997). *Methodological approaches to online scoring of essays* (Technical Report 461). UCLA. National Center for Research on Evaluation, Student Standards, and Testing. Los Angeles: University of California.

Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment, 5*(1), 1-35. Retrieved [date] from http://www.jtla.org.

Dikli, S., & Bleyle, S. (2014). Automated essay scoring feedback for second language writers: How does it compare to instructor feedback?. *Assessing Writing, 22,* 1-17.

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

Elliot, S. (2001). *Applying IntelliMetric technology to the scoring of 3rd and 8th grade standardized writing assessments* (RB-524). Newtown, PA: Vantage Learning.

Enright, M., & Quinlan, M. (2010). Complementing human judgment of essays written by English language learners with e-rater® scoring. *Language Testing, 27*(3), 317-334.

Ericsson, P. F., & Haswell, R. H. (2006). *Machine scoring of student essays: Truth and consequences*. Logan, Utah: Utah State University Press.

Fang, Y. (2010). Perceptions of the computer-assisted writing program among EFL college learners. *Journal of Educational Technology & Society, 13*(3), 246-256.

Foltz, P. W., Kintsch W., & Landauer, T. K. (1998). The measurement of textual coherence with Latent Semantic Analysis. *Organizational Process, 25*(2-3), 285-307.

Grimes, D., & Warschauer, M. (2010). Utility in a fallible tool: A multi-site case study of automated writing evaluation. *Journal of Technology, Language, and Assessment, 8*(6), 1-43.

Hamp-Lyons, L. (1990). Second language writing: Assessment issues. In B. Kroll (Ed.), *Second language writing* (pp. 69-87). United Kingdom: Cambridge University Press.

Hamp-Lyons, L. (2001). English for academic purposes. In R. Carter & D. Nunan (Eds.), *The Cambridge TESOL guide* (pp. 126-130). Cambridge: Cambridge University Press.

Hoang, G. T. L., & Kunnan, A. J. (2016). Automated essay evaluation for English language learners: A case study of MY Access. *Language Assessment Quarterly, 13*(4), 359-376.

Homburg, T.J. (1984). Holistic evaluation of ESL composition: Can it be validated objectively?. *TESOL Quarterly, 18*(1), 87-108.

Horning, A. (1987). *Teaching writing as a second language*. United States of America: SIU Press.

Huang, J. (2008). How accurate are ESL students' holistic writing scores on large-scale assessments? - A generalizability theory approach. *Assessing Writing, 13*(3), 201-218.

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

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

Hyland, K. (2003). Writing and teaching writing. In Richards, J. C. (Eds.), *Second language writing* (pp. 1-28). United States of America: Cambridge University Press.

Hyland, F., Nicolas-Conesa, F., & Cerezo, L. (2016). Key issues of debate about feedback on writing. In R. M. Manchón & P. K. Matsuda (Eds.), *Handbook of second and foreign language writing* (pp. 433-452). Berlin, Germany: Walter de Gruyter GmbH.

James, C. L. (2006). Validating a computerized scoring system for assessing writing and placing students in composition courses. *Assessing Writing, 11*, 167-178.

Johnson, R.L., Penny, J.A., & Gordon, B. (2009). *Assessing performance: Designing, scoring, and validating performance tasks*. New York: The Guilford Press.

Kellogg, R. T., Whiteford, A. P., & Quinlan, T. (2010). Does automated feedback help students learn to write?. *Journal of Educational Computing Research, 42*(2), 173-196.

Lai, Y. H. (2010). Which do students prefer to evaluate their essays: Peers or computer program. *British Journal of Educational Technology, 41*(3), 432-454.

Landauer, T. K., Foltz, P. W., & Laham, D. (1997). An introduction to latent semantic analysis. *Discourse Processes, 25*(2-3), 259-284.

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

Li, Z., Link, S., Ma, H., Yang, H., & Hegelheimer, V. (2015). The role of automated writing evaluation holistic scores in the ESL classroom. *System, 44*, 66-78.

Liao, H. C. (2015). Using automated writing evaluation to reduce grammar errors in writing. *ELT Journal, 70*(3), 308-319.

Link, S., Dursun, A., Karakaya, K., & Hegelheimer, V. (2014). Towards better ESL practices for implementing automated writing evaluation. *Calico Journal, 31*(3), n3.

Liu, S., & Kunnan, A. J. (2016). Investigating the application of automated writing evaluation to Chinese undergraduate English majors: A case study of "WriteToLearn". *Calico Journal, 33*(1), 71-91.

Maeng, U. (2010). The effect and teachers' perception of using an automated essay scoring system in L2 writing. *English Language and Linguistics, 16*(1), 247-275.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13-103).New York: Macmillan.

McNamara, T. F. (1996). *Measuring second language performance*. London and New York, NY: Addison Wesley Longman.

Myers, M. (2003). What can computers and AES contribute to a K–12 writing program? In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 3-20). Mahwah, NJ: Lawrence Erlbaum.

Page, E. B. (1967). The imminence of grading essays by computer. *Phi Delta Kappan, 48*, 238-243.

Parr, J. M., & Timperley, H. S. (2010). Feedback to writing, assessment for teaching and learning and student progress. *Assessing Writing, 15*(2), 68-85.

Petersen, N. S. (1997). *Automated scoring of writing essays: Can such scores be valid.* In annual meeting of the National Council on Education, Chicago, IL.

Popham, J.W. (1981). *Modern educational measurement*. Englewood: Prentice.

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

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2002b). Stumping e-rater: challenging the validity of automated essay scoring. *Computers in Human Behavior, 18*(2), 103-134.

Rock, J. (2007). *The impact of short-term use of Criterion on writing skills in ninth grade* (Research Report 07-07). Princeton, NJ: Educational Testing Service.

Shermis, M. D., & Burstein, J. (2003). *Automated Essay Scoring: A cross disciplinary perspective*. Mahwah, NJ: Lawrence Erlbaum.

Shermis, M., Burstein, J., Higgins, D., & Zechner, K. (2010). Automated essay scoring: Writing assessment and instruction. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International encyclopedia of education* (3rd ed., pp. 20-26). Oxford, UK: Elsevier.

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

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.

Spencer, B., & Louw, H. (2008). A practice-based evaluation of an on-line writing evaluation system: First-World technology in a Third-World teaching context. *Language Matters, 39*(1), 111-125.

Steinhart, D. J. (2001). *Summary Street: An intelligent tutoring system for improving student writing through the use of Latent Semantic Analysis* (Unpublished doctoral dissertation). University of Colorado, Boulder.

Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. *Assessing Writing, 19*, 51-65.

Tang, J., & Rich, C. S. (2017). Automated writing evaluation in an EFL setting: Lessons from China. *JALT CALL Journal, 13*(2), 117-146.

Tsuda, N. (2014). Implementing Criterion (Automated Writing Evaluation) in Japanese college EFL classes. *Language and Culture: The Journal of the Institute for Language and Culture, 18*, 25-45.

Wang, P. (2015). Effects of an automated writing evaluation program: Student experiences and perceptions. *Electronic Journal of Foreign Language Teaching, 12*(1), 79–100.

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

Wang, Y. J., Shang, H. F., & Briody, P. (2013). Exploring the impact of using automated writing evaluation in English as a foreign language university students' writing. *Computer Assisted Language Learning, 26*(3), 234-257.

Warschauer, M., & Grimes, D. (2008). Automated writing assessment in the classroom. *Pedagogies, 3*(1), 52-67.

Warschauer, M., & Ware, P. (2006*).* Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research, 10*(2), 1-24.

Weigle, S. C. (2013). English as a second language writing and automated essay evaluation. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 58-76). New York and London: Routledge.

Williamson, D., Xi, X., & Breyer, J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement, Issues and Practice, 31*(1), 2-13.

Wilson, J., & Czik, A. (2016). Automated essay evaluation software in English Language Arts classrooms: Effects on teacher feedback, student motivation, and writing quality. *Computers & Education, 100*, 94-109.

Yagelski, R. P. (1995). The role of classroom context in the revision strategies of student writers. *Research in the Teaching of English*, 216-238.

Elif Sari *is an EFL instructor at the School of Foreign Languages at Karadeniz Technical University, Turkey. She earned her Ph.D. degree at Atatürk University in 2020. Her research areas include writing assessment with a special focus on automated feedback and scoring, motivation, and grammar teaching in the EFL context.*

Email: elifsari@ktu.edu.tr

Turgay Han *is an Associate Professor and Head of English Language and Literature Department at Ordu University. His research areas include L2 measurement and assessment issues with a special focus on individual differences in language learning and generalizability (G-) theory, the examination of affective factors in L2, the use of mobile applications in L2 learning, and the application of different assessment frameworks in assessing L2 writing performance.*

Email: turgayhan@odu.edu.tr