

RESEARCH ARTICLE

Voluntary use of automated writing evaluation by content course students

Aysel Saricaoglu

TED University, Turkey (saricaogluaysel@gmail.com)

Zeynep Bilki

TED University, Turkey (zeynep.bilki@tedu.edu.tr)

Abstract

Automated writing evaluation (AWE) technologies are common supplementary tools for helping students improve their language accuracy using automated feedback. In most existing studies, AWE has been implemented as a class activity or an assignment requirement in English or academic writing classes. The potential of AWE as a voluntary language learning tool is unknown. This study reports on the voluntary use of Criterion by English as a foreign language students in two content courses for two assignments. We investigated (a) to what extent students used Criterion and (b) to what extent their revisions based on automated feedback increased the accuracy of their writing from the first submitted draft to the last in both assignments. We analyzed students' performance summary reports from Criterion using descriptive statistics and non-parametric statistical tests. The findings showed that not all students used Criterion or resubmitted a revised draft. However, the findings also showed that engagement with automated feedback significantly reduced users' errors from the first draft to the last in 11 error categories in total for the two assignments.

Keywords: automated writing evaluation; voluntary use; accuracy improvement; content course students

1. Introduction

One of the major effects of globalization on education is the increasing number of universities in non-English speaking countries that adopt English as the medium of instruction (EMI) (Dang, Nguyen & Le, 2013). In EMI contexts, in addition to learning content knowledge, students are expected to develop their language competences (Wilkinson, 2013). However, the content staff in EMI programs “rarely see their role as one of developing the students' language ability” (Wilkinson, 2013: 16). Because instructors often employ a content-oriented focus in EMI, students' language needs may not be equally addressed. Such an educational landscape highlights the need for students to confront the challenge of continuous language improvement on their own. Fortunately, today's technologies such as automated writing evaluation (AWE) have the potential to assist them with this challenge.

AWE tools evaluate students' written language in seconds and offer specific and individualized feedback (Chapelle & Voss, 2016; Chun, 2016; Wang, Shang & Briody, 2013). The benefits of automated feedback from AWE tools for students, including writing longer essays, obtaining higher machine scores, making fewer language errors, and improving the rhetorical quality of writing, are readily acknowledged in the literature (Feng, 2015; Wang, 2013; Wang *et al.*, 2013). Most existing AWE studies were conducted in English or writing classes where the use

Cite this article: Saricaoglu, A. & Bilki, Z. (2021). Voluntary use of automated writing evaluation by content course students. *ReCALL* 33(3): 265–277. <https://doi.org/10.1017/S0958344021000021>

© The Author(s), 2021. Published by Cambridge University Press on behalf of European Association for Computer Assisted Language Learning. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

of an AWE tool was a requirement for a certain assignment and was led by the instructor (e.g. Chen & Cheng, 2008; Lavolette, Polio & Kahng, 2015; Saricaoglu, 2019). Today, more AWE tools are offered to students as self-led supplementary aids, but the potential of AWE for such use is still unknown, which is the concern of the current study. Addressing this gap, we aimed to explore English as a foreign language (EFL) students' voluntary use of Criterion, an AWE tool developed by Educational Testing Service. More specifically, we are concerned with students' engagement with automated feedback from Criterion for their content course assignments and their error reduction rates as a result of making use of automated feedback.

1.1 Students' engagement with automated feedback

Researchers examining students' engagement with AWE feedback have generally looked at details of the revision process, such as whether students revised their written drafts based on automated feedback, how much time they spent revising a draft, how many resubmissions they made, whether their revisions were correct or incorrect, what type of revisions they made (i.e. correction, substitution, addition, or deletion), or whether students used any strategies (i.e. cognitive or metacognitive) in the revision process (e.g. Lavolette *et al.*, 2015; Zhang & Hyland, 2018).

Zhang and Hyland (2018) describe a "highly engaged learner" as the one who makes efforts in the revision task to improve the quality of their writing based on the automated feedback received from Pigai, an AWE tool working in a similar way to commonly known AWE systems such as Criterion. The student who had a high level of engagement with Pigai resubmitted her essay to the AWE system 13 times and increased her score as a result from 79 to 90. Another characteristic of the highly engaged learner was that she did not complete the revisions immediately but over two weeks, having some gaps between her resubmissions. She was also emotionally invested in improving her essay and was motivated to see her score increase after revisions. On the other hand, Zhang and Hyland (2018) describe a "moderately engaged learner" as the one who does not make enough effort in the revision task to improve their writing. For example, the student who had a moderate level of engagement with automated feedback resubmitted his essay to Pigai only once and did not address several of the errors identified automatically. He also reported to be overwhelmed by the feedback that he received.

In the investigation of students' engagement with automated feedback, factors that influence students' AWE use are also important. The findings of Grimes and Warschauer (2010) suggest that the amount of revision using an AWE tool increases in time as teachers allow more time for writing assignments. In the first year of their MY Access! implementation, only 12% of student essays were revised. In the third year, this percentage increased to 53. Lavolette *et al.* (2015) investigated the effects of three factors on students' revisions using Criterion: feedback category, the accuracy of error identification, and the timing of the feedback. Their findings showed that students made related revisions for the ill-formed verb, proofread, and subject-verb agreement errors about 85% of the time and about 50% of the time for the missing comma, preposition, and spelling errors, but they ignored capitalization errors 41% of the time. They revised incorrectly identified errors about half the time. Whether the feedback was immediate or delayed did not have a significant effect on their revisions. In a study on AWE feedback on students' written scientific arguments, Zhu *et al.* (2017) found that 77% of the students revised their writing after receiving automated feedback. They also found that those whose initial AWE scores were higher were more likely to revise their writing. According to Ranalli (2018), how specific or generic the feedback is determines students' abilities to correct their errors, with specific feedback resulting in more successful corrections. He found course level not to have a significant effect on students' error correction abilities when using Criterion. In a survey study, Li *et al.* (2019) found that learners' intention to use Pigai was influenced by perceived usefulness, attitude towards using, computer self-efficacy, and perceived ease of use.

Research on AWE has also highlighted the critical role of teacher and/or peer feedback that follows automated feedback. In Lai's (2010) study, students adopted peer feedback more often than

automated feedback from MY Access! as they found automated feedback to be too general and indirect, whereas peer feedback was direct and specific. Chen and Cheng (2008) compared three ways of integrating MY Access! into EFL college-level writing courses. In the first integration, students had to achieve a minimum score of 4 out of 6 from the system. They also received written teacher feedback and in-class peer feedback. In the second integration, students used the tool as much as they needed without a requirement of a certain score. They also received written teacher feedback but no peer feedback. In the third integration, students used the program autonomously with the instructor being less involved. They did not receive any teacher feedback, only online peer feedback. Chen and Cheng (2008) found that AWE feedback was perceived more positively by the students when it was followed by teacher and peer feedback. In fact, using the AWE tool without human facilitation made students frustrated. These findings are supported by Wilson and Czika (2016) who explored students' writing motivation in two conditions: automated feedback from PEG Writing, developed by Measurement Incorporated, combined with teacher feedback condition and only electronic teacher feedback condition. Teacher feedback combined with automated feedback was found to increase students' motivation and to lead to a higher number of essay drafts, but no statistically significant gains were observed in writing quality between the two conditions.

1.2 Automated feedback and error reductions

Previous studies on the effects of AWE feedback on the accuracy of students' writing have reported mixed but generally positive findings. Based on the increase in the automated holistic scores that students received from their first essays to their fourth, El Ebyary and Wendeatt (2010) concluded that the accuracy of students' writing improved as a result of making use of automated feedback from Criterion. Wang *et al.* (2013) compared the accuracy of students' writing before and after they used CorrectEnglish. Their findings revealed significantly fewer errors in run-on sentences, sentence fragments, capitalization errors, missing articles, and punctuation in the post-test. Lavolette *et al.* (2015) compared the essays of students who received immediate (40 minutes after submission) and delayed (three weeks after submission) feedback from Criterion and observed no significant difference between the feedback groups. There was also no difference between the groups in accuracy over time. Liao (2016), who investigated whether using Criterion improved the accuracy of students' writing in revised texts and in new texts, found that AWE feedback had a positive effect on the reduction of errors both in text revision and in text composition. Students' accuracy significantly improved in four error categories: fragments, subject-verb disagreement, run-on sentences, and ill-formed verbs. Li, Feng and Saricaoglu (2017) examined the short-term (from the first draft to the last draft of the same paper) and the long-term (from the first draft of a paper to the first draft of a subsequent paper) effects of Criterion feedback on ESL students' development of grammatical accuracy. The findings regarding the short-term effects showed that students reduced errors in eight error categories: word choice, verb form, articles, pronouns, run-on sentences, fragments, sentence structure, and subject-verb agreement. In the long term, there was significant error reduction in only one error category: run-on sentences.

Our study adds to the AWE literature reviewed above by focusing on the voluntary use of Criterion by content course students. It investigates students' engagement with Criterion as a voluntary language learning tool and the improvement in the accuracy of their written assignments as a result of using automated feedback. The study was guided by the following research questions:

1. To what extent do students voluntarily use Criterion for their content course assignments?
2. To what extent does students' voluntary use of Criterion increase the accuracy of their writing?

2. Methodology

This study was conducted as an action research project since it emerged as a context-specific effort to find a solution to an educational problem: students' lack of language support in content courses. Action research, involving both action and research activities, intervenes in social processes in particular contexts, such as classrooms, schools, or organizations, for positive change (Burns, 2009). It starts with the identification of a problem or question, continues with the planning and implementation of an action, and ends with observation and reflection on the outcome and future plans (Lewin, 1946). We had prior knowledge about and experiences of Criterion and brought our belief to the inquiry process that students would benefit from using Criterion for their content course assignments. We took a quantitative approach to investigate students' voluntary use of Criterion and the improvement in the accuracy of their writing as a result of engaging with automated feedback.

2.1 Context and participants

This study was conducted at a private Turkish university with English as the medium of instruction. Volunteer sampling was used to recruit the participants. All faculty members were sent an email introducing the study, and the members who believed that their students would benefit from using Criterion for their course assignments were invited to the study. Faculty members were informed that this implementation would not create any extra workload for them, and that all the work, including training the students, creating course assignments on Criterion, and tracking students' use of the program, would be handled by the researchers.

The number of Criterion seats were limited to 200 student accounts. Thus, the researchers met face to face with the first six faculty members who expressed an interest and provided them with more details about the study. Three of the faculty members were teaching two sections of the same course, and one faculty member was teaching three sections of the same course. Table 1 presents the distribution of courses. In total, students from five different courses and 10 classes taught by six faculty members were involved in the study: Materials Science (three sections), Research Project, Introduction to Education (two sections), Community Service (two sections), and Introduction to Sociology (two sections). Courses, assignment genres, and the number of assignments across courses are presented in Table 1.

The medium of instruction in all classes was English, and excellence in English was one of the priorities of the university. All faculty members were native speakers of Turkish, but they were all proficient speakers of English. The faculty members teaching the Materials Science, Introduction to Sociology, Introduction to Education, and Community Service courses were assistant professors. The faculty member teaching the Research Project course was a full professor.

Out of 199 students in 10 classes, 114 (female = 81, male = 33) volunteered to participate in this study. Table 2 displays the number of volunteers from each class. The language proficiency level of students when they start their departmental programs is upper-intermediate (B2) according to the Common European Framework of Reference for Languages (Council of Europe, 2001). They either submit an international English language proficiency exam score (84 TOEFL IBT or 6.5 IELTS Academic) or pass the English language proficiency exam administered by the university with a minimum score of 80, which is again an equivalent of 84 TOEFL IBT or 6.5 IELTS Academic.

2.2 AWE tool

Criterion was chosen as the AWE tool in this study as researchers had prior classroom experience with it. Criterion is an online application that offers students several writing-specific features, including planning, holistic scores, trait-specific feedback on different types of errors, unlimited revision and resubmission opportunities, as well as peer and instructor feedback opportunities.

Table 1. Courses, assignment genres, and assignment numbers

Course	Year	Assignment genre	Assignment number
Materials Science	2nd	Lab report	4
Research Project	4th	Research article	1
Introduction to Education	1st	Written reflection	5
		Opinion essay	1
		Guest speaker report	3
Community Service	4th	Project process journal	1
		Project proposal	1
Introduction to Sociology	1st	Movie review	2

Table 2. Courses, years, number of students, and number of volunteers

Course	Year	Number of students	Number of volunteers
Materials Science	2nd	39	23
Research Project	4th	9	0
Introduction to Education	1st	51	37
Community Service	4th	48	19
Introduction to Sociology	1st	52	35
	Total	199	114

Different from other AWE applications such as Grammarly, which not only allows the user to upload a completed text for instant evaluation but also provides corrective feedback during the process of text creation, Criterion provides instant feedback on a written text once the completed text is submitted to the system. One advantage that Criterion has over other AWE applications is that it logs user data and generates reports (e.g. score analysis reports or detailed or summary performance reports) or trait feedback analysis, including number of words or sentences and numbers of errors in specific error categories that could be used for instructional or research purposes.

Criterion generates feedback in five categories: organization and development, grammar, usage, mechanics, and style. Because Criterion gives organization and development feedback for the essay genre, we disabled this feedback category, and students only received feedback on grammar, usage, mechanics, and style. Table 1A in the supplementary material presents Criterion's error categories and specific types of errors in each category. Criterion does not display student errors from different categories all together on a written text. In order to see feedback from each error category, the user needs to click on a specific type of error from a specific error category, as displayed in Figure 1. The user then needs to move the cursor to the highlighted error in order to see the feedback on that specific error, as seen in Figure 2.

Although Criterion has a rich library of essay topics, we used the customized topics since students wrote in different genres such as lab reports, research articles, reflections, journals, projects proposals, and movie reviews, the topics of which were determined by the faculty members.

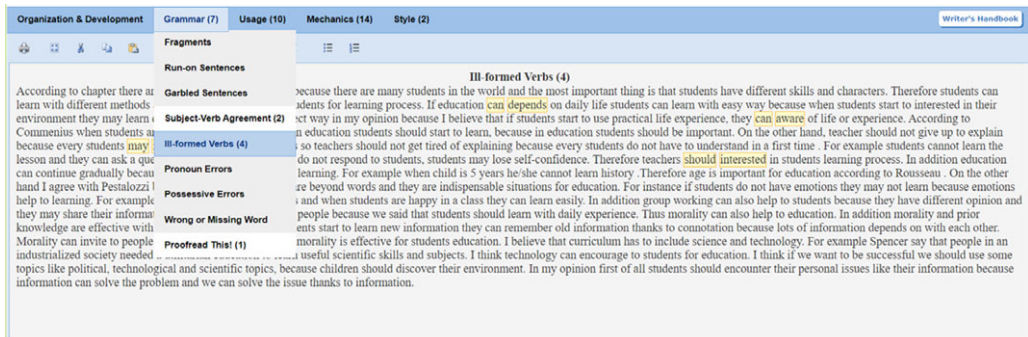


Figure 1. Criterion's display of identified errors

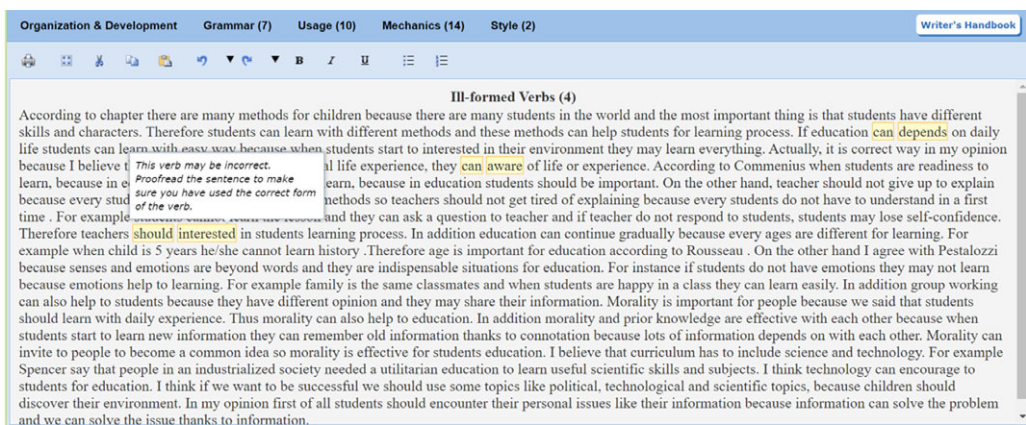


Figure 2. Criterion's feedback on a highlighted ill-formed verb error

2.3 Procedures

Classes of the six faculty members were visited to introduce the study to the students, train them, and recruit volunteers. Before class visits, student accounts were created on Criterion for all students taking the selected courses. During the training, students were informed that they all had access to Criterion and were provided with their student account details. Students were trained in how to log in to the system, how to open an assignment, how to submit a text for evaluation, how to see the automated feedback, how to revise language errors, and how to resubmit the text. They were delivered a guidebook including details of how to use Criterion. After a 15-minute training, they were informed about the study and were invited as participants. It was explicitly stated that they would be able to use Criterion for their class assignments whether or not they volunteered to share their AWE data. Consent forms were delivered and signed by those volunteering.

During the semester, content course assignments were created on Criterion by the researchers. Students were informed via email that the tool was ready for their use. Although faculty members were not asked to do anything specific for this project, some faculty members expressed several times during our conversations that they orally encouraged their students to use Criterion. Faculty members teaching the Materials Science, Research Project, and Introduction to Education courses stated that they offered frequent oral encouragement in their courses to facilitate their students' Criterion use. Faculty members teaching the Introduction to Sociology classes decided to offer

their students two bonus points in each assignment for using Criterion and reducing their errors. For this reason, these two faculty members were given a report of their students' Criterion use including students' number of submissions and error counts after each submission.

2.4 Data collection and analysis

Although five courses were involved in the study, we collected data from only two courses. We excluded the Materials Science course from the data collection process because students worked in groups for each assignment. It would not be possible to track individual students' use of the tool. We also excluded the Community Service course as there was only a very small number of volunteers. We did not collect data from the Research Project course as there were no volunteers. We collected data only from the Introduction to Education (IE) and Introduction to Sociology (IS) courses. Both were first year courses with a similar number of volunteers. While IE students used Criterion for nine assignments, IS students used it for only two assignments. Therefore, we used data from only the first two assignments in both courses.

In order to explore students' Criterion use and error responses, we collected expanded performance summary reports from Criterion for each volunteer and each assignment. The summary reports included information regarding an individual student's resubmission numbers and number of errors in specific error categories in each submitted draft. We transferred the submission numbers and error counts to an Excel sheet for statistical analyses. To address the effect of text length on the number of errors students made, we normalized the raw error counts per 1,000 words. Because data were not normally distributed, we used the Wilcoxon signed-rank test (the non-parametric dependent samples *t*-test) to compare error reduction rates from students' first submitted drafts to the last submitted drafts within the same assignments.

3. Results and discussion

3.1 Students' voluntary use of Criterion for content course assignments

To answer the first research question, we examined the number of Criterion users (i.e. number of students who made at least one submission to Criterion) in each assignment and their number of resubmissions (i.e. how many times they resubmitted the same draft to Criterion after revising their language errors based on the automated feedback) in each assignment. Table 3 presents the descriptive findings.

In the first assignment, 46% of IE students and 71% of IS students used Criterion. In both courses, some students did not submit any revised drafts (three IE students and five IS students). In the second assignment, fewer IE students (32%) used Criterion than in the first assignment, but they made more resubmissions ($M = 3.36$) in the second assignment than in the first assignment ($M = 2.79$). On the other hand, there was no decrease in the number of Criterion users from the IS course in the second assignment, and there was a slight increase in the number of their resubmissions ($M = 3.10$ in the first assignment and $M = 3.57$ in the second assignment). Again, not all students revised and resubmitted their drafts in the second assignment. Four students from both courses made only one submission and did not submit a revised draft.

In contrast to previous AWE implementations where AWE use was a class activity or an assignment requirement, the decision of whether or not to use Criterion was made by the students in this study. As a result, the participation rate of students in AWE use was much lower than the participation rate of students in those earlier studies (both for the number of Criterion users and the number of resubmissions). When AWE is used during the class or when a teacher asks students to use it (e.g. Lavolette *et al.*, 2015; Li *et al.*, 2017; Saricaoglu, 2019; Wang *et al.*, 2013), almost all students use the AWE tool to meet the requirement. Explicit integration of

Table 3. Number of Criterion users and resubmissions across classes and assignments

Course	Assign	N	Users		Resubmitters		Resubmissions	
			#	%	#	%	M	SD
Introduction to Education	1st	37	17	46%	14	82%	2.79	1.19
Introduction to Sociology	1st	35	25	71%	20	80%	3.10	1.37
Introduction to Education	2nd	37	12	32%	8	22%	3.36	2.44
Introduction to Sociology	2nd	35	25	71%	21	60%	3.57	1.21

Note. N = number of volunteers.

AWE technology in the learning process in such implementations may better promote its use by students, as Koh (2016) argues.

Students' lower rates of AWE use in this study might also be related to the lack of teacher involvement. In one study, Chen and Cheng (2008) revealed that without teacher facilitation, the autonomous use of MY Access! was frustrating for students. In another study, Lai, Yeung and Hu (2016) pointed out that students needed greater support from their teachers in contrast to the minimal responsibility that teachers assumed in promoting out-of-class language learning with technology. In our study, although the teachers facilitated students' use of Criterion through oral encouragement or extra points, students did not receive any explicit teacher and/or researcher support, for instance, in terms of how to interpret the feedback, how to address the feedback, whether to revise all identified errors or not, and so on, which may have negatively impacted the extent to which they used Criterion.

We noticed that more IS students used Criterion in both assignments than IE students. Also, a higher percentage of IS students resubmitted their drafts in the second assignment than IE students. These findings could be related to how course instructors promoted using Criterion. While the IE instructor explained that they provided frequent oral encouragement to facilitate their students' Criterion use, IS instructors offered their students two bonus points for using Criterion in each assignment. Bonus points might have been more encouraging to the students. However, since no qualitative follow-up data were collected as to the reasons why some students did not use Criterion or why they did not resubmit a draft, it is hard to interpret the findings.

3.2 Increase in the accuracy of students' writing

To answer the second research question, we compared students' first and last submitted drafts from the same assignment for error counts. Table 4 presents the descriptive findings of students' error reduction rates in Assignment 1. The highest reduction rate of IE students belongs to usage errors (43%) followed by grammar errors (41%) and mechanics errors (28%). The lowest error reduction rate was in the style category (21%). IS students reduced their grammar errors (59%) the most in Assignment 1, followed by usage errors (41%) and mechanics errors (30%). They also had the lowest error reduction rate in the style category (16%).

Table 5 presents the descriptive findings of students' error reduction rates from the first submission to the last submission in Assignment 2. Both groups had the highest reduction rate (59%) in grammar errors (75% for IE students and 37% for IS students), followed by usage errors (45% for IE students and 30% for IS students), and mechanics errors (27% for IE students and 20% for IS students). Again, style errors had the lowest reduction rate (8% for both groups).

Whether there was a significant improvement in the accuracy of students' drafts in one assignment was measured using the Wilcoxon signed-rank test. Error counts in the first submission were compared with the error counts in the last submission of the same assignment. According to the findings presented in Table 2A in the supplementary material, in Assignment 1, IE students

Table 4. Error reduction rates in Assignment 1 across classes

Class	Error category	N	First draft			Last draft			Error reduction rates ^a
			#	<i>M</i>	<i>SD</i>	#	<i>M</i>	<i>SD</i>	
Introduction to Education	Grammar	14	167	11.83	10.75	98	7.00	8.61	41%
	Usage	14	255	18.07	8.09	146	10.43	10.90	43%
	Mechanics	14	422	30.21	43.38	304	21.71	35.67	28%
	Style	14	459	32.86	43.85	361	25.79	41.67	21%
Introduction to Sociology	Grammar	20	199	10.00	7.44	81	4.05	4.10	59%
	Usage	20	405	20.40	11.77	239	11.95	12.23	41%
	Mechanics	20	332	16.65	9.98	232	11.60	10.14	30%
	Style	20	341	17.05	24.14	288	14.40	24.17	16%

Note. N = text number; # = normalized error counts.

^aError reduction rates show the percentage of reduced errors in the last submission made by the students in an error category compared with the number of errors in the first submission.

Table 5. Error reduction rates in Assignment 2 across classes

Class	Error category	N	First draft			Last draft			Error reduction rates ^a
			#	<i>M</i>	<i>SD</i>	#	<i>M</i>	<i>SD</i>	
Introduction to Education	Grammar	8	71	9.88	5.28	18	2.25	2.38	75%
	Usage	8	121	16.75	9.74	67	10.38	7.05	45%
	Mechanics	8	245	34.50	20.10	180	23.50	21.07	27%
	Style	8	278	36.38	34.77	255	33.38	35.01	8%
Introduction to Sociology	Grammar	21	208	9.95	8.45	130	6.19	8.25	37%
	Usage	21	423	20.00	9.24	295	14.05	11.48	30%
	Mechanics	21	653	31.24	13.99	521	24.81	13.84	20%
	Style	21	572	27.14	31.18	529	25.19	31.37	8%

Note. N = text number; # = normalized error counts.

^aError reduction rates show the percentage of reduced errors in the last submission made by the students in an error category compared with the number of errors in the first submission.

made significant error reductions in Subject-Verb Agreement ($z = -2.56$, $p = .011$, with a large effect size, $r = -.68$), Possessive ($z = -2.04$, $p = .041$, with a large effect size, $r = -.55$), Missing Article ($z = -2.56$, $p = .011$, with a large effect size, $r = -.68$), and Missing Comma ($z = -2.03$, $p = .042$, with a medium effect size, $r = -.54$). In Assignment 2, their error rates significantly decreased for Subject-Verb Agreement ($z = -2.03$, $p = .042$, with a large effect size, $r = -.72$) and Missing Comma ($z = -2.38$, $p = .017$, with a large effect size, $r = -.84$).

According to the findings presented in Table 3A in the supplementary material, in Assignment 1, IS students' error reductions from the first submission to the last were significant in the following error types: Garbled Sentences ($z = -2.33$, $p = .020$, with a large effect size, $r = -.52$), Subject-Verb Agreement ($z = -2.99$, $p = .003$, with a large effect size, $r = -.67$), Possessive ($z = -1.93$, $p = .054$, with a medium effect size, $r = -.43$), Proofread This! ($z = -2.70$, $p = .007$, with a large effect size, $r = -.60$), Missing Article ($z = -2.48$, $p = .004$, with a large effect size, $r = -.55$), Confused Words ($z = -2.85$, $p = .004$, with a large effect size, $r = -.64$), Wrong Form of Word ($z = -2.27$, $p = .023$, with a large effect size, $r = .51$), Missing Comma ($z = -2.27$, $p = .023$, with a large effect size, $r = -.51$), and Extra Comma ($z = -2.39$, $p = .017$, with a large effect size, $r = -.53$).

In Assignment 2, students significantly reduced their Subject-Verb Agreement errors ($z = -2.40$, $p = .016$, with a large effect size, $r = -.54$), Missing Article errors ($z = -2.91$, $p = .004$, with a large effect size, $r = -.65$), Spelling errors ($z = -2.03$, $p = .042$, with a medium effect size, $r = -.45$), Missing Initial Capital Letter in a Sentence errors ($z = -1.93$, $p = .054$, with a medium effect size, $r = -.43$), and Missing Comma errors ($z = -2.25$, $p = .024$, with a large effect size, $r = -.50$).

In summary, students from both classes significantly reduced their errors in the following specific error types: Garbled Sentences, Subject-Verb Agreement, Possessive, Missing Article, Confused Words, Wrong Form of Word, Proofread This!, Missing Comma, Extra Comma, Spelling, and Missing Initial Capital Letter in a Sentence. These findings are consistent with findings from earlier research on AWE, which have already reported the positive impact of automated feedback on the accuracy of students' writing (e.g. Li, Link & Hegelheimer, 2015; Liao, 2016; Wang *et al.*, 2013). The findings are also in line with theoretical perspectives from second language acquisition. From the interactionist perspective, feedback draws the learner's attention to language forms during L2 interaction, which has a major role in language learning (Gass & Mackey, 2006; Long 1996; Polio, 2012). "Written correction is a form of feedback that gives learners an indication of their errors" (Polio, 2012: 383). Applying the interactionist perspective to AWE, automated feedback indicates students' errors during L2 interaction that occurs between the students and an AWE tool, thus draws their attention to language forms and positively affects their learning.

While the number of students' errors in certain categories significantly decreased from their first submission to the last within the same assignment, such an effect is described as a short-term effect in the AWE literature (Li *et al.*, 2017). The long-term effects of automated feedback on students' error reductions (i.e. from their first submitted draft of the first assignment to their first submitted draft of the last assignment) were not investigated in this study as we only examined the first two assignments from both groups and not the same students within one group used Criterion for the two assignments. Given that previous research has yielded limited or no benefits of AWE feedback in the long term (Lavolette *et al.*, 2015; Saricaoglu, 2019), we would not expect long-term benefits in voluntary use of AWE either. However, with an interactionist perspective, we consider short-term changes as evidence of learning (Mackey & Polio, 2009; Norris & Ortega, 2003; Polio, 2012). Moreover, whether the benefits of corrective feedback are long term or short term, as stated by Ferris, Liu, Sinha and Senna (2013), teachers continue responding to students' language errors, thus the important question is how to provide it in the best way. Today, with automated feedback supporting teacher or peer feedback in writing practices, the same question also applies to automated feedback practices. To this end, development of effective strategies for using automated feedback seems to be one important goal for writing instructors and researchers.

4. Conclusion

English remains a challenge for content course students because the focus is generally on the subject matter rather than language (Leki, 2006). Existing technologies such as AWE can create language learning opportunities for students in such contexts (Plonsky & Ziegler, 2016). In this study, we attempted to explore the potential of AWE as a voluntary language learning tool for content course assignments. We specifically examined EFL students' voluntary use of Criterion for content course assignments and the improvement in the accuracy of their writing as a result of making use of automated feedback. Using Criterion was not an assignment requirement, but an individual decision of students. Our results showed that not all students used Criterion or attempted to revise their drafts using automated feedback. However, for those who submitted revised drafts, engagement with automated feedback significantly reduced their language errors from the first draft to the last in 11 error categories in total for the two assignments.

Findings of this study should be considered in light of several limitations. First, the present study did not control possible confounds that might have influenced students' voluntary use

of AWE, such as their error correction backgrounds, motivation levels, assignment assessment criteria, and so on. No teacher data were obtained regarding how much feedback they gave to students' English language in their written assignments, and whether the quality of their language was one of the assignment evaluation criteria or not. What importance students gave to their English language in their assignments would possibly be influenced by what importance their teachers gave to language, which as a result could affect students' AWE use. Researchers of future studies might usefully include measuring such possible factors.

Second, no qualitative data were collected in this study. Without qualitative data that can shed light on students' engagement with Criterion, several questions remain unanswered. Why did some students make only one submission in an assignment and did not attempt to correct their errors? We cannot know if this was because they did not understand the feedback or because they did not know how to correct their errors. Did the provision of oral encouragement or bonus points have a role in students' engagement with Criterion? We cannot know the extent to which oral encouragement or bonus points could have served as a factor underlying students' Criterion use and revision of errors. The lack of qualitative follow-up data creates a limitation in understanding such matters. Researchers of future studies will need to include a qualitative inquiry into students' AWE use to avoid such limitations.

Third, in this study, automated feedback on organization was disabled as our focus was on language errors and improvement across drafts. Because the organization feedback was disabled, we also disabled the scoring function of the tool. Thus, we do not know if there would be differences in students' Criterion scores across drafts. Zhu *et al.* (2017) showed that students who corrected their errors using automated feedback received significantly higher final scores than those who did not and found an association between each revision and an average increase of 0.55 on the scores. Referring to students being "score-obsessed," Zhang and Hyland (2018: 96) highlight the motivating effect of seeing an improvement in the scores for students. Score increases themselves also serve as a type of feedback to students confirming their error corrections and their language improvement. Thus, a useful area for future research would be to confirm the improvement in students' accuracy with the increase in automated scores along with its motivating effect for more revisions.

Another limitation of the study is that we did not explore how accurate Criterion's error identification was, even though our second research question focused on the increase in accuracy across students' drafts. The capacity of automated systems' error identification compared to humans' error identification is a controversy that is still debated in the AWE literature (e.g. Hoang & Kunnan, 2016; Koltovskaia, 2020). Studies by Dikli and Bleyle (2014), Lavolette *et al.* (2015), and Ranalli, Link and Chukharev-Hudilainen (2017) have shown that Criterion's error identification accuracy level is much lower than the determined thresholds of 80% or 90% (Burstein, Chodorow & Leacock, 2003; Quinlan, Higgins & Wolff, 2009). Given that Criterion operates at low error identification capacity, the lack of human raters and the lack of human-generated feedback add to the limitations of our study. Yet because students used Criterion in their content courses where they did not receive human feedback on their language errors, we believe that receiving Criterion feedback was more beneficial than receiving no feedback at all.

Finally, the analysis of the quality of the learners' texts in this study was limited to the number of errors detected by Criterion across drafts. Evaluating text quality from the perspective of the teachers would also have been helpful given that the study was based on a small corpus of texts. A qualitative or quantitative exploration of whether teachers thought Criterion users submitted noticeably better assignments in terms of language would have provided a better understanding of the effect of automated feedback. Future research should benefit from additional measures of text quality when investigating the effect of automated feedback in voluntary use of AWE.

Supplementary material. To view supplementary material referred to in this article, please visit <https://doi.org/10.1017/S0958344021000021>

Acknowledgements. This study was supported by an institutional research grant from TED University. We would like to thank Burak Şenel who worked as a research assistant in the project. We presented this study at EuroCALL 2018 Conference, and we thank the audience for their valuable comments.

Ethical statement. The authors reported no potential conflict of interest.

References

- Burns, A. (2009) Action research in second language teacher education. In Burns, A. & Richards, J. C. (eds.), *The Cambridge guide to second language teacher education*. New York: Cambridge University Press, 289–297.
- Burstein, J., Chodorow, M. & Leacock, C. (2003) *Criterion*SM online essay evaluation: An application for automated evaluation of student essays. In Riedl, J. & Hill, R. (eds.), *Proceedings of the Fifteenth Innovative Applications of Artificial Intelligence Conference*. Menlo Park: AAAI Press, 3–10.
- Chapelle, C. A. & Voss, E. (2016) 20 years of technology and language assessment in language learning & technology. *Language Learning & Technology*, 20(2): 116–128. <https://www.lltjournal.org/item/2950>
- Chen, C.-F. E. & Cheng, W.-Y. E. (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. <https://www.lltjournal.org/item/2631>
- Chun, D. M. (2016) The role of technology in SLA research. *Language Learning & Technology*, 20(2): 98–115. <https://www.lltjournal.org/item/2949>
- Council of Europe (2001) *Common European framework of reference for languages: Learning, teaching, assessment*. Strasbourg: Language Policy Unit.
- Dang, T. K. A., Nguyen, H. T. M. & Le, T. T. T. (2013) The impacts of globalisation on EFL teacher education through English as a medium of instruction: An example from Vietnam. *Current Issues in Language Planning*, 14(1): 52–72. <https://doi.org/10.1080/14664208.2013.780321>
- 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. <https://doi.org/10.1016/j.asw.2014.03.006>
- 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. <https://files.eric.ed.gov/fulltext/EJ936915.pdf> <https://doi.org/10.6018/ijes/2010/2/119231>
- Feng, H.-H. (2015) *Designing, implementing, and evaluating an automated writing evaluation tool for improving EFL graduate students' abstract writing: A case in Taiwan*. Iowa State University, unpublished PhD. <https://lib.dr.iastate.edu/etd/14824/>
- Ferris, D. R., Liu, H., Sinha, A. & Senna, M. (2013) Written corrective feedback for individual L2 writers. *Journal of Second Language Writing*, 22(3): 307–329. <https://doi.org/10.1016/j.jslw.2012.09.009>
- Gass, S. M. & Mackey, A. (2006) Input, interaction and output: An overview. *AILA Review*, 19(1): 3–17. <https://doi.org/10.1075/aila.19.03gas>
- Grimes, D. & Warschauer, M. (2010) Utility in a fallible tool: A multi-site case study of automated writing evaluation. *Journal of Technology, Learning, and Assessment*, 8(6): 4–43.
- 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. <https://doi.org/10.1080/15434303.2016.1230121>
- Koh, C. (2016) Translating motivational theory into application of information technology in the classroom. In Liu, W. C., Wang, J. C. K. & Ryan, R. M. (eds.), *Building autonomous learners: Perspectives from research and practice using self-determination theory*. New York: Springer, 245–258. https://doi.org/10.1007/978-981-287-630-0_13
- Koltovskaia, S. (2020) Student engagement with automated written corrective feedback (AWCF) provided by *Grammarly*: A multiple case study. *Assessing Writing*, 44: 1–12. <https://doi.org/10.1016/j.asw.2020.100450>
- Lai, C., Yeung, Y. & Hu, J. (2016) University student and teacher perceptions of teacher roles in promoting autonomous language learning with technology outside the classroom. *Computer Assisted Language Learning*, 29(4): 703–723. <https://doi.org/10.1080/09588221.2015.1016441>
- 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. <https://doi.org/10.1111/j.1467-8535.2009.00959.x>
- Lavolette, E., Polio, C. & Kahng, J. (2015) The accuracy of computer-assisted feedback and students' responses to it. *Language Learning & Technology*, 19(2): 50–68. <https://www.lltjournal.org/item/2903>
- Leki, I. (2006) "You cannot ignore": L2 graduate students' response to discipline-based written feedback. In Hyland, K. & Hyland, F. (eds.), *Feedback in second language writing: Contexts and issues*. New York: Cambridge University Press, 266–285. <https://doi.org/10.1017/CBO9781139524742.016>
- Lewin, K. (1946) Action research and minority problems. *Journal of Social Issues*, 2(4): 34–46. <https://doi.org/10.1111/j.1540-4560.1946.tb02295.x>
- Li, J., Link, S. & Hegelheimer, V. (2015) Rethinking the role of automated writing evaluation (AWE) feedback in ESL writing instruction. *Journal of Second Language Writing*, 27: 1–18. <https://doi.org/10.1016/j.jslw.2014.10.004>


- Li, R., Meng, Z., Tian, M., Zhang, Z., Ni, C. & Xiao, W. (2019) Examining EFL learners' individual antecedents on the adoption of automated writing evaluation in China. *Computer Assisted Language Learning*, 32(7): 784–804. <https://doi.org/10.1080/09588221.2018.1540433>
- Li, Z., Feng, H.-H. & Saricaoglu, A. (2017) The short-term and long-term effects of AWE feedback on ESL students' development of grammatical accuracy. *CALICO Journal*, 34(3): 355–375. <https://journals.equinoxpub.com/index.php/CALICO/article/view/26382>
- Liao, H.-C. (2016) Using automated writing evaluation to reduce grammar errors in writing. *ELT Journal*, 70(3): 308–319. <https://doi.org/10.1093/elt/ccv058>
- Long, M. H. (1996) The role of the linguistic environment in second language acquisition. In Ritchie, W. C. & Bhatia, T. K. (eds.), *Handbook of second language acquisition*. San Diego: Academic Press, 413–468. <https://doi.org/10.1016/B978-012589042-7/50015-3>
- Mackey, A. & Polio, C. (2009) Introduction. In Mackey, A. & Polio, C. (eds.), *Multiple perspectives on interaction: Second language research in honor of Susan M. Gass*. New York: Routledge, 1–10. <https://doi.org/10.4324/9780203880852>
- Norris, J. & Ortega, L. (2003) Defining and measuring SLA. In Doughty, C. J. & Long, M. H. (eds.), *Handbook of second language acquisition*. Malden: Blackwell, 717–761.
- Plonsky, L. & Ziegler, N. (2016) The CALL–SLA interface: Insights from a second-order synthesis. *Language Learning & Technology*, 20(2): 17–37. <https://www.lltjournal.org/item/2945>
- Polio, C. (2012) The relevance of second language acquisition theory to the written error correction debate. *Journal of Second Language Writing*, 21(4): 375–389. <https://doi.org/10.1016/j.jslw.2012.09.004>
- Quinlan, T., Higgins, D. & Wolff, S. (2009) *Evaluating the construct-coverage of the e-rater scoring engine*. Princeton: Educational Testing Service. <https://doi.org/10.1002/j.2333-8504.2009.tb02158.x>
- Ranalli, J. (2018) Automated written corrective feedback: How well can students make use of it? *Computer Assisted Language Learning*, 31(7): 653–674. <https://doi.org/10.1080/09588221.2018.1428994>
- Ranalli, J., Link, S. & Chukharev-Hudilainen, E. (2017) Automated writing evaluation for formative assessment of second language writing: Investigating the accuracy and usefulness of feedback as part of argument-based validation. *Educational Psychology*, 37(1): 8–25. <https://doi.org/10.1080/01443410.2015.1136407>
- Saricaoglu, A. (2019) The impact of automated feedback on L2 learners' written causal explanations. *ReCALL*, 31(2): 189–203. <https://doi.org/10.1017/S095834401800006X>
- Wang, P.-I. (2013) Can automated writing evaluation programs help students improve their English writing? *International Journal of Applied Linguistics & English Literature*, 2(1): 6–12. <https://doi.org/10.7575/ijale.v.2n.1p.6>
- 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. <https://doi.org/10.1080/09588221.2012.655300>
- Wilkinson, R. (2013) English-medium instruction at a Dutch university: Challenges and pitfalls. In Doiz, A., Lasagabaster, D. & Sierra, J. M. (eds.), *English medium instruction at universities: Global challenges*. Tonawanda: Multilingual Matters, 3–24. <https://doi.org/10.21832/9781847698162-005>
- 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. <https://doi.org/10.1016/j.compedu.2016.05.004>
- Zhang, Z. V. & Hyland, K. (2018) Student engagement with teacher and automated feedback on L2 writing. *Assessing Writing*, 36: 90–102. <https://doi.org/10.1016/j.asw.2018.02.004>
- Zhu, M., Lee, H.-S., Wang, T., Liu, O. L., Belur, V. & Pallant, A. (2017) Investigating the impact of automated feedback on students' scientific argumentation. *International Journal of Science Education*, 39(12): 1648–1668. <https://doi.org/10.1080/09500693.2017.1347303>

About the authors

Aysel Saricaoglu is an assistant professor in English language and literature at Social Sciences University of Ankara, Turkey. Her research involves academic writing, automated writing evaluation, syntactic complexity, and technology-enhanced language learning. She has disseminated her work in journals such as *Studies in Second Language Acquisition*, *Assessing Writing*, *Computer Assisted Language Learning*, and *ReCALL*.

Zeynep Bilki is an assistant professor in the English language education program at TED University, Turkey. Her research interests include second language reading and writing, technology-enhanced foreign language learning, and language teacher education. She has been an English language educator in Turkey and USA.

Author ORCID.  Aysel Saricaoglu, <https://orcid.org/0000-0002-5315-018X>

Author ORCID.  Zeynep Bilki, <https://orcid.org/0000-0001-9505-8093>