Modern English Education

http://dx.doi.org/10.18095/meeso.2019.20.4.18

Engagement With Automated Writing Feedback in Mandated vs. Voluntary Conditions*

Shaun J. Manning**
Hankuk University of Foreign Studies

Sookyung Cho
Hankuk University of Foreign Studies

ARTICLE INFO

Received: 15 September 2019
Revised: 21 October 2019
Accepted: 8 November 2019

Examples in: English
Applicable Languages: English
Applicable Levels: Tertiary

KEYWORDS

EFL Writing/
AWE/
content feedback/
error correction /
student perceptions/
EFL 쓰기/
자동채점/
내용 피드백/
오류 수정/
학생의 인식

ABSTRACT

Manning, S. J., & Cho, Sookyung. (2019). Engagement with automated writing feedback in mandated vs. voluntary conditions. *Modern English Education, 20*(4), 18-30.

This study investigates student engagement with Écree, a Natural Language Processing (NLP) based automated writing evaluation (AWE) that gives feedback on content. In particular, this study examines how mandating AWE use or not, affects students' behavioral and emotional engagement with AWE feedback. Two classes—one in which Écree use was mandatory (EM) and another in which it was optional (EO)—were surveyed and four volunteers from each class were tracked. Overall, the two groups differed in their evaluations of how much they incorporated AWE feedback into revisions as well as in the number of uploads to Écree. Tracking the focal students' written drafts, Écree feedback, and interview data revealed that students from EM tended to incorporate Écree's feedback into their revisions more than EO students. Unexpectedly, EM students developed quite negative views toward Écree, whereas EO students often referred to its usefulness: its systematicity, speed, and role as a reader. This study implies that AWE can be utilized as a supplemental tool in higher education and that students may well need training so as to maximize its positive effects.

I. INTRODUCTION

Since automated writing evaluation (AWE) emerged with the development of Page Essay Grade in 1960, it has developed at an amazing pace and has been widely used in various contexts. Despite many studies that buttress the validity and reliability of AWE compared to human raters, Warschauer and Grimes (2008) claim the use of AWE in a writing classroom remains "in the midway between the fears of some and the hopes of others" (p. 22). On one

* This paper was supported by Hankuk University of Foreign Studies Research Fund of 2019.
** First author: Shaun J. Manning, Corresponding author: Sookyung Cho

Shaun J. Manning (Professor)
Department of English Linguistics and Language Technology, Hankuk University of Foreign Studies, 107 Imun-ro, Dongdaemun-gu, Seoul, 02450, Korea
Tel: (02) 2173-2266 / Email: shaunmanning@yahoo.com

Sookyung Cho (Professor)
Department of English Linguistics and Language Technology, Hankuk University of Foreign Studies, 107 Imun-ro, Dongdaemun-gu, Seoul, 02450, Korea
Tel: (02) 2173-3194 / Email: sookyungcho@hufs.ac.kr

hand, proponents of AWE argue that it assists both the instructors and students (Burstein, Chodorow, & Leacock, 2004; Elliot & Mikulas, 2004; Foltz, Laham, & Landauer, 1999; Grimes & Warschauer, 2010; Myers, 2003); AWE eases the writing instructors' burden of marking and fixing errors and thus enables them to focus on the content; and in the case of students, AWE's feedback and scores motivate them to write and revise more. On the other hand, skeptics claim that AWE thwarts the role of writing instructors and distorts students' views of good writing (Baron, 1998; CCCC Executive Committee, 2004; Cheville, 2004; Ericsson & Haswell, 2006).

Aside from the issue of whether to use AWE or not, yet other scholars emphasize that we must consider how to use AWE in the classroom (Chen & Cheng, 2008; Williamson, 2004). Given the necessity to see how to utilize AWE in a writing class, it has been argued that it is a pity that, to date, the majority of studies have focused on the effects of AWE on students' test scores in standardized writing tests (Chen & Cheng, 2008; Warshauer & Ware, 2006). Most studies have compared a control group who did not use an AWE with an experimental group who used one and then highlighted the positive effects of AWE by associating the experimental group's improvement in their score and grammatical accuracy with the AWE (Attali, 2004; Attali & Burstein, 2006; Shermis, Burstein, & Bliss, 2004; Shermis, Garvan, & Diao, 2008; Vantage Learning, 2007). By looking at the numerical values such as increased scores in writing, without seeing the actual usage of AWE in a writing classroom, however, we are not able to unveil any pedagogical implications of AWE (Li, Link, & Hegelheimer, 2015).

Out of such necessity to see the usefulness of AWE in a writing class, a few studies have recently begun to explore how writing instructors integrate AWE into their classes and how students use its feedback. Some studies investigated writing instructors' and students' use of AWE (Grimes & Warschauer, 2010; Wang, Shang, & Briody, 2013; Warschauer & Grimes, 2008) while others examined students' attitudes towards the use of AWE in writing courses (Chen & Cheng, 2008; Lai, 2010; Li et al., 2015). These studies seem to agree that the usefulness of AWE depends on a wide variety of variables such as teachers' perspective on AWE (Grimes & Warschauer, 2010; Li et al., 2015; Warschauer & Grimes, 2008), students' proficiency (Li et al., 2015), stages of writing process (Chen & Cheng, 2008), or types of feedback (Chen & Cheng, 2008; Li et al., 2015). Drawn upon the assumption that the manner in which an AWE is implemented in the classroom matters when investigating an AWE's effects, this study investigates another potential way that an AWE might be used by teachers: mandating its use or leaving it up to each student to decide whether or not to use it. By looking at student revisions, perceptions, and attitudes towards the feedback they received from Écree, this study asks: Does requiring the use of this AWE (Écree) affect students' behavioral and emotional engagement with its feedback?

II. LITERATURE REVIEW

Research on AWE can be loosely divided into studies that attempt to prove that using the AWE improved writing on a variety of dimensions, and research into how AWE were implemented by instructors and used by students. We look at each approach in turn.

1. The Effectiveness Approach

Since the introduction of the first type of AWE, Essay Grade Page, in 1960, vendors and scholars who had great interest in promoting AWE have put great effort into proving the benefits of AWE to fight skepticism against it raised by some writing scholars and instructors (Attali, 2004; Shermis et al., 2004; Shermis, Garvan, & Diao, 2008; Vantage Learning, 2007). For example, Attali (2004) investigated the effects of *Criterion*—a web-based system of automated writing evaluation, developed by ETS—based on 9,725 essays written by students from sixth to 12th grades in the USA. After comparing their first submission of an essay with the last one, Attali found that students' revised essays improved not only in terms of grammar, but also in discourse elements by using more background and conclusion elements, main points, and supporting ideas. Similarly, Shermis et al. (2004) and Shermis et al. (2008) confirmed the positive effects of AWE on student essays by investigating the *Criterion* feedback on a large number of essays in the USA. After comparing a group of tenth graders who used *Criterion* with a control group who did not, Shermis et al. (2004) concluded that those who used *Criterion* wrote more, scored higher, and made fewer errors than their counterparts. Expanding the age of students who used *Criterion* to sixth to eighth and tenth grades, Shermis et al. (2008) investigated 11,685 essays written by 2,017 students to seven different prompts and found overall improvement in production variables, such as essay score, essay length, and number of unique words, as well as in writing errors—grammar, usage, mechanics, style, organization and development. In particular, they noticed a peak improvement at the eighth grade.

Interestingly, although the above studies found that AWE promoted improved writing, one attempt to evaluate an AWE's performance with EFL learners found it to be inferior to human raters. J. Park (2019) working with Korean high school students' writing found that a very popular, AI based grammar checker – *Grammarly Premium* missed a large number of simple errors that human raters detected, including verb agreement, verb tense, and word choice errors. She concluded that grammar checkers can be helpful but that users need to "accept the limitations of the AI-based grammar checker" (J. Park, 2019, p. 129) implying that human judgement was also needed for successful use of AWE.

2. The Implementation Approach

While the previously mentioned studies examined the effects of AWE on scores or on the performance of the AEW vis-à-vis humans, other studies have addressed the issue of how AWE affects writing instruction by exploring the use of AWE in actual writing classrooms (Grimes & Warschauer, 2010; Wang, Shang, & Briody, 2013; Warschauer & Grimes, 2008). Warshauer and Grimes (2008) conducted a mixed-methods exploratory case study on one middle school (Grades 6 to 8), two junior high schools (Grades 7 to 8), and one high school (Grades 9 to 12). After interviewing both the teachers and the students, Warschauer and Grimes argued that both teachers and students had a limited use of AWE, while having positive attitudes towards it. This analysis of student essays based on feedback of various AWE services—*Criterion*, *My Access*, and *WriteToLearn*—reveals that AWE led to a greater amount of revision, but only on a superficial level. After conducting a multi-site case study of AWE on eight middle schools in Southern California over a three-year period, Grimes and Warschauer (2010) also found some positive effects of AWE both for teachers and students. That is, AWE simplifies classroom management by making teachers more relaxed when students worked with AWE, and it motivates students to write and revise more.

However, it seems that the success of AWE in assisting a writing course depends on various factors. According to Chen and Cheng (2008), stages of writing process and availability of additional feedback other than AWE itself seem to affect the effects of AWE. After investigating the use of *My Access!* in three different classroom settings in Taiwan, Chen and Cheng found that AWE can be utilized most and is highly appreciated by the students under the following circumstances: when AWE is used in the early stages of drafting and when it is accompanied by human assistance such as teacher and peer feedback.

Additionally, Li et al. (2015) argue that students' proficiency and writing experience affect their attitudes toward the use of AWE in a writing class. They found that students who did not have a lot of academic writing experience were satisfied with AWE feedback, in particular, on the organization of their essays, while those who had relatively more experience with academic writing were not satisfied with AWE feedback.

Moreover, the teachers seem to play an important role in the ways in which AWE is useful to the students (Grimes & Warschauer, 2010; Li et al., 2015; Warschauer & Grimes, 2008). For instance, in Li et al. (2015), students' preference of AWE corrective feedback was based on their instructors' views and perspectives on AWE, for their teachers emphasized the value of its corrective feedback and utilized AWE as a tool that provides corrective feedback in their writing courses.

As these studies show, we need to consider how to effectively use AWE in our writing courses rather than whether to use it or not (Chen & Cheng, 2008; Williamson, 2004). Based on the same assumption that the mode of using AWE matters when investigating its effects, this study investigates whether to require the use of AWE or not: how a mandate to use AWE affects students' engagement with AWE.

Several studies compare and contrast the effects of mandatory versus optional use of a computer in various areas of education (Armel & Shrock, 1996; Brandl, 2012; Garland & Noyes, 2004). However, their findings on the nature of computer use—whether to use it as a requirement or optional—are inconclusive. Armel and Shrock (1996) and Garland and Noyes (2004) favored the required use of the computer, online note-taking and a computer-based learning package respectively, whereas Brandl (2012) found different strengths of each group—the optional group produced more output while the required group produced more accurate output. None of these studies, however, concerns the use of AWE in the classroom. In order to fill this gap, this study tries to examine how the nature of AWE use, such as whether to use it as mandatory or optional affects learners' writing as in this study.

Moreover, this study focuses on content feedback provided by an AWE system named Écree (Heit & Donaldson, 2018). So far, most studies have focused on formal aspects of writing such as language and discourse. Most studies on AWE feedback have been using *Criterion* (Attali, 2004; Li et al., 2015; Shermis et al., 2004; Shermis et al., 2008) or *My Access!* (Chen & Cheng, 2008; Grimes & Warschauer, 2010; Wang et al., 2013; Warschauer & Grimes, 2008; Vantage Learning, 2007). Although both AWE systems provide feedback on areas other than grammar, such as content, development, or organization, they are mainly focused on grammar. For example, *Criterion*, which consists of two complimentary applications, *e-rater* and *Critique*, diagnoses writers' uses of grammar and style based on their errors in grammar, usage, mechanics, style, organization segments, and vocabulary content. Similarly, *My Access!* rates an essay and gives feedback based on the five areas—focus and meaning, content and development, organization, language use and style, and mechanics and conventions. As Li et al. (2015) argue, however, it is necessary to separate feedback on grammar from feedback on content in order to see the effects of AWE more precisely. By using Écree, a program specialized in providing feedback on contents (see Section III. 2), this study fits into this niche and investigates the role of automated writing feedback on contents in the context of EFL writing instruction and its effect on student engagement, focusing on their behaviors and emotions. The research questions of this study are as follows:

1) Does mandating make a difference in students' overall engagement with Écree?
2) How does mandating affect students' behavioral engagement with Écree?
3) How does mandating affect students' emotional engagement with Écree?

TABLE 1
Participant Information

Group	Student	Gender	Year in university	Estimated English level CEFR (iBT TOEFL*)	English experience
Optional	Sumie	F	4	C1 (95)	Learned English only in Korea, except one semester study in the USA as a university exchange student. Never attended an after-school academy.
	Minho	M	4	B2+ (72)	Has never studied abroad. Only learned English in Korea, both in school and at academies.
	Jinhee	F	3	C1+ (95)	Studied in the USA as an undergraduate student before returning to Korea to re-do university. Teaches English at an academy.
	Semin	F	2	B2+ (72)	Has never studied abroad. Only learned English in Korea, both in school and at academies.
Mandatory	Yejin	F	4	B1+ (42)	Has never studied abroad. Only learned English in Korea. Teaches English to middle school students at an academy.
	Wonjoon	M	4	C1+ (95)	Lived in the USA for over three years as an elementary and middle school student.
	Mina	F	2	C1 (95)	Has never studied abroad. Only learned English in Korea, both in school and at academies.
	Youngmee	F	1	B1+ (42)	Has never studied abroad. Only learned English in Korea, both in school and at academies.

* English level is estimated based on a students' essay written during the second class of the semester. The iBT TOEFL equivalents are cut-lines (minimum scores on the TOEFL iBT needed to achieve that CEFR level) based on Papageorgiou, S., Tannenbaum, R. J., Bridgeman, B., & Cho, Y. (2015) and are included for the convenience of those unfamiliar with CEFR levels. The students' actual TOEFL scores were not known – none of the participants reported having taken a TOEFL recently.

III. METHOD

This study adopts a mixed-method approach: a quantitative analysis of two classes—Écree-mandatory (EM) and Écree-optional (EO)—and an in-depth analysis of focal students from each class.

1. Participants

The students were enrolled in an undergraduate writing course, English Writing 1, intended for students majoring in English at a university in Seoul, South Korea. There were two sections of English Writing 1, each meeting once per week for two hours. The EM section had 21 students (14 female, 7 male) while the EO section had 18 students (11 female, 7 male). Both sections were taught by the same instructor, one of the authors of this study. He is a North-American male with a doctorate in Applied Linguistics and over 20 years' experience teaching EFL writing in South Korea. Both sections used the same materials, were assigned the same essay assignments, and did the same in-class activities. The key difference between the two classes was that the EM class was required to upload their assignments to the AWE (Écree) prior to submitting them for grading, while the EO class had the AWE available to them but was not required to upload. To ensure the EM class uploaded, the instructor asked for a printout of the Écree feedback to be brought to the class during the peer review session and for a printout to be submitted with the final draft of the paper.

For the in-depth analysis, four focal students were recruited from each section (eight volunteers total). Each volunteer was a native speaker of Korean with a high proficiency of English. All the focal students had taken university English writing classes prior to the study. Sumie and Yejin had studied writing previously with the same instructor. So, all had been exposed to the rhetorical and linguistic demands of English writing at the university level; i.e., they were learning to write but were not novice writers. Table 1 lists the detailed demographic information of each focal student.

2. What is Écree?

Écree is an AWE based on natural language processing (NLP) designed for university writing programs in the United States. Écree is set up by instructors for use with classes; it is not available for individual student use. The instructor sets up a class, and students register for Écree, and join the class. Once the class is set up, the instructor inputs the assignments for that class. Prior to the study the instructor had used Turnitin (www.turnitin.com), Grammarly (www.grammarly.com), and the Virtual Writing Tutor (virtualwritingtutor.com) both for his own writing and with his students. He had never used Écree prior to this semester but chose to use it because he felt its advantages of different genres, feedback on content, and teacher control over the final score would be helpful.

A key distinguishing feature of Écree is the availability of different feedback criteria for different genres of essay—including argumentative essays, process essays, single paragraphs, example essays, cause-effect essays, etc. This feature mitigates the criticism that learners may focus only on one type of writing (e.g., the five-paragraph essay), because that is what typical AWE grade. When inputting an assignment, the instructor selects the genre from a drop-down menu, and inputs some keywords that the algorithm will use when grading. The students will then be graded according to criteria Écree has established for that genre. Écree will score the writing within minutes of an upload and give feedback to the student. However, Écree does not give a score to the student—that goes to the instructor who is free to assign the score as is, adapt the score based on their criteria for the class, or ignore

the score completely. The instructor is not committed to using the AWE-assigned score.

Because students do not see a number, they are expected to engage with written commentary feedback which advises them what to change. The student sees a screen with their writing separated into paragraphs, each paragraph evaluated separately along several dimensions. The student's writing is at the top, with feedback on the AWE's criteria, including: length, topic sentence, supporting evidence, analysis, and level of detail below it. Criteria which the AWE assesses as being successful are shaded in green, less successful criteria are yellow, while criteria that need more attention are shaded in red. It is up to the student to determine what the commentary is really asking for, or to ask a peer or the instructor for assistance.

3. Data Collection

This study took place over a 16-week semester in the spring of 2017. Various types of data were collected from all the participants, eight focal students, and the instructor.

1) Survey and Number of Uploads to Écree

At the end of the semester all students were surveyed (about Écree, and some general questions). There were three types of questions: (a) disagree or agree, which students answered on a Likert scale of 1-6, 1 = strongly disagree and 6 = strongly agree; (b) a question about the frequency of an occurrence, which they also used a 1-6 Likert Scale from 1 = never to 6 = always; and (c) open ended questions. Both the disagree-agree and frequency questions also had space for the students to write in their own thoughts about the question if they wished (see Appendix 1 for the questions regarding Écree).

The number of uploads per student per assignment was retrieved from Écree which stores all student uploads until the course is deleted. The total number of (re)uploads and all alterations between submissions could be tracked and compared across EM and EO groups.

2) Student Essay Drafts and Interviews

From the eight focal students, all of their drafts were collected along with feedback provided by Écree on each draft. Additionally, they were interviewed two times, in the 7th and 13th weeks of the semester. The second author of this study, who is a native speaker of Korean, interviewed the participants. We felt that having an outsider as opposed to the class instructor conduct the interviews and do so in the students' first language provided two advantages. First, the students might feel freer to criticize the instructor, Écree, or any other aspects of the course. We thought that if students had something negative to say, they may hesitate to say it directly to their instructor, but would feel more comfortable talking with a different person who could protect their identity. Second, although some of the volunteers' English level was advanced (see Table 1), using their first language would let them discuss their feelings and ideas more deeply and precisely than if they used English. During the interview, they were mostly asked about their experiences with Écree feedback, including its usefulness when they worked on revisions and their overall evaluations of it. These interviews were semi-structured, and lasted approximately 15 minutes each; they were audio-recorded. (See Appendix 2 for the English version of the interview questions).

3) Classroom Behavior: Instructor Observation, Reflection, and Audio Recording

As the lead researcher was also the teacher of the classes, his observations in class, his post-lesson reflections, his reading of the Écree feedback to the students, along with comments students made to him form part of the data set relevant to the study.

The classroom observation was conducted in the following manner. Each lesson he created a seating chart after students had been put into small groups, and he would place an audio recorder at each group. As it was the instructor's normal class procedure to evaluate student participation (which comprised 20% of the final score) based on their talk, the students were familiar with the recorders. During group work as he moved around the class, he would note any groups and times which had any critical incidents, problems, or interesting comments that he felt required listening to. Listening to these key points allowed him the opportunity to reflect on the lesson in general (Farrell & Kennedy, 2019) and in particular, on how the students were reacting to Écree. The audio recordings also provide a record of student reaction to Écree.

4. Data Analysis

The first research question was answered by the comparison and contrast of the survey data between EO and EM groups, using Wilcoxon signed rank tests since the small sample cannot be assumed to be normally distributed. Also, the number of submissions to Écree was compared between the two groups. With regard to the second research question, students' behavioral engagement with Écree, their incorporation of Écree feedback into the revisions was analyzed. All the Écree feedback on students' each draft was identified and their incorporations into the next drafts were checked. The number of words involved in each revision was counted and averaged by the number of drafts because of the necessity to control the variation in the amount of words—because the more drafts are submitted, the more likely it is that the number of words involved in revision will increase. Finally, to examine the students' emotional engagement with Écree, all the recordings of the participants' interviews were transcribed word by word and analyzed based on the inductive approach suggested by Leki (2006). First, the transcribed data were reviewed repeatedly until potential themes or categories emerged out of the data. Later, they were tabulated against the audiotapes of the data in order to identify additional dimensions or aspects which could be easily lost when only the written scripts were examined. Once those emerging themes

and categories were identified, they were read again and tabulated against the research questions and, if relevant, compared and contrasted with the instructor's observations and reflections.

IV. RESULTS

1. Does Mandating Make a Difference in Students' Overall Engagement With Écree?

1) Analysis of Survey Data

Overall, the survey was found to have Cronbach's Alpha, $\alpha = 0.768$, suggesting an acceptable level of reliability for Likert data. Table 2 shows results of the survey conducted on both EM and EO classes. A series of Wilcoxon signed rank tests reveal that only one question—Question 14 (I was able to fully apply Écree's feedback to revisions of my essay)—marked a significant difference between Mandatory and Optional groups. The Mandatory group's average response was 2.77 (out of 6) while the Optional group gave an average score of 3.64. This indicates stronger disagreement with the statement from the Mandatory group—students who were required to use the AWE reported being less capable of using its suggestions in their revisions. A boxplot of the student responses to this question is given in Figure 1.

TABLE 2
Results of Wilcoxon Signed Rank Tests for Écree Survey

Part 2: Écree	Average (mean)			Wilcoxon tests	
Question	EM ($n = 22$)	EO ($n = 11^{\dagger}$)	Difference	*W*	*p*
13 (D or A) I revised the parts of my essay which Écree did not classify as green.	3.50	4.00	-0.50	96.50	0.34
14 (D or A) I was able to fully apply Écree's feedback to revisions of my essay.	2.77	3.64	-0.86	66.00	0.03*
15 (D or A)I trusted Écree's feedback about:					
(a) Language (Grammar and spelling)	4.43	4.55	-0.12	116.50	0.97
(b) Organization (Thesis statements, etc.)	3.95	4.09	-0.14	114.00	0.97
(c) Content (Examples, details, ideas)	3.43	3.82	-0.39	94.50	0.40
18 How often did you disagree with Écree feedback?	3.64	3.55	0.09	132.00	0.67
20 (D or A) The feedback from Écree was sometimes not relevant to my point.	4.00	3.45	0.55	154	0.19

† The optional group had 17 students, but six students never used Écree, so their data has been removed from the quantitative analysis.
* = statistically significant result
** = close to statistical significance

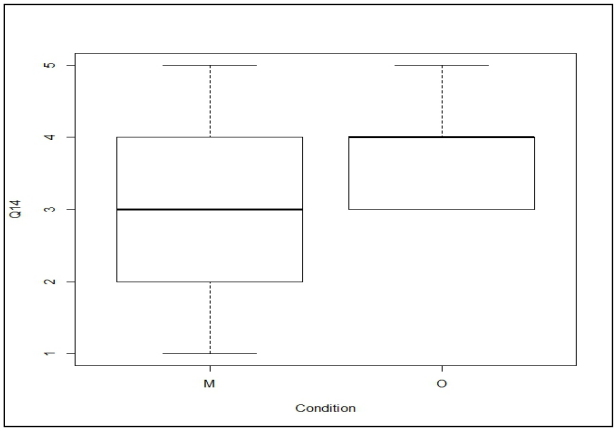

FIGURE 1 Boxplot of Student Responses Regarding Ability to Apply Écree Feedback in Later Writing

The boxplot shows that there were no 1's or 2's assigned by students in the optional use class, whereas some mandatory students gave the lower scores. The lack of variation may be the result of self-selection bias, only students who felt they could use Écree's feedback may have opted to use Écree.

2) Number of Uploads to Écree

Table 3 shows the total number of uploads by all the participants. Two trends are apparent: (1) students in the EM condition uploaded far more than the EO condition, and (2) Écree use decreased as the semester progressed.

TABLE 3
Number of Uploads for Each Assignment in Each Condition

Condition	Assignment 1 Argumentative essay	Assignment 2 Example essay	Assignment 3 Cause-effect essay	Total (per condition)	Average per student (*SD*)
EM	134	90	60	284	12.91 (13.35)
EO	27	21	13	61	3.39 (5.56)
Total (*M, SD*)	161 (4.13, 5.81)	111 (2.85, 4.14)	73 (1.88, 2.31)	345	

Students in the EM condition uploaded over four times more frequently than their EO peers; however, in both classes, the total number of uploads was less than half for the third assignment than for the first one. The total difference in uploads was statistically significant with a large effect size ($t = 3.0382$, $df = 29.237$, $p = 0.004973$, Cohen's $d = 0.93$). This simply shows that those who were required to upload uploaded more.

In addition, there was a decrease in total uploads from 161 in the first assignment to 73 in the final assignment. The results of paired samples tests show that this decrease was also statistically significant, ($t = 3.172$, $df = 39$, $p = 0.002948$, Cohen's $d = 0.51$). This demonstrates that students used the AWE less frequently as the course progressed, either because they felt more confident in their writing, or they felt less confident in the AWE. To determine this, we examined the Écree use by eight focal students who volunteered to participate in interviews.

TABLE 4
Incorporation of Écree Feedback

Group	Student	Writing assignment						Mean per submission
		1st		2nd		3rd		
		Submission (words)	Ratio	Submission (words)	Ratio	Submission (words)	Ratio	
EO	Sumie	0(0)	**0.00**	0(0)	**0.00**	0(0)	**0.00**	0.00
	Minho	0(0)	**0.00**	1(0)	**0.00**	1(0)	**0.00**	0.00
	Jinhee	2(172)	**86.00**	5(338)	**67.60**	2(144)	**72.00**	72.67
	Semin	3(328)	**109.33**	2(491)	**245.50**	1(111)	**111.00**	155.00
EM	Yejin	3(909)	**303.00**	2(20)	**10.00**	3(156)	**520.00**	312.13
	Wonjoon	3(1016)	**338.67**	2(424)	**212.00**	2(343)	**171.50**	254.71
	Mina	2(71)	**35.50**	3(346)	**115.33**	2(92)	**46.00**	72.71
	Youngmee	7(571)	**81.57**	2(21)	**10.50**	3(18)	**6.00**	50.83

2. How Does Mandating Affect Students' Behavioral Engagement With Écree?

As the number of uploads has shown that EM group uploaded their drafts on Écree more frequently than EO group, it is worth investigating whether they really reflected Écree feedback in their revisions or not.

Table 4 shows the amount of incorporation of Écree feedback into each revision between EO and EM classes. The EM class seems to have a tendency to incorporate more Écree feedback than the EO, in contrast to EM class' low-level of appreciation of Écree feedback found in the survey. The four focal participants from EM made revisions based on Écree feedback, although there existed individual variations in the extent of revision, whereas the two participants in EO did not incorporate Écree feedback at all in their revision. Because Sumie did not use Écree across the semester, naturally none of her changes in the revision originated from Écree feedback; Minho tried Écree once each for his second and third writing assignment but did not incorporate its feedback at all in his revision. It is quite interesting to see there is a mismatch between his remarks on Écree feedback during the interview and the actual results. At the interview (see below), Minho mentioned that Écree feedback was helpful when he checked whether to provide enough details or not, but in reality, he submitted the same draft without any changes, although Écree suggested that he needed more details in his introduction and the first paragraph.

The analysis of student drafts uncovers another interesting pattern, which is the decrease in the extent to which most participants incorporated Écree feedback across the writing assignments, whether they were in EM or EO. This finding corresponds with what has been found in the number of uploads, which also decreased through the semester. Except for Yejin in the mandatory class (from 303 in the 1st assignment to 520 in the 2nd one), the other participants adopted Écree feedback comparatively less in the last writing assignments than in their prior writing assignments. Semin from the optional group and Mina from the mandatory class peaked in their second writing assignment but decreased uploading in their third assignment; both Wonjoon and Youngmee in the mandatory class gradually decreased in their changes drawn upon Écree feedback across the writing assignments. Jinhee in the optional group made a few more changes based on Écree feedback in her last writing assignment than in her second one, but still made fewer changes than she did in her first writing assignment. This decreased use of Écree feedback may relate to their perceptions on Écree feedback: during the interviews (see below), both Wonjoon and Youngmee strongly complained about inconsistency of Écree feedback. Their distrust of Écree as a reliable source of providing feedback may have led them to incorporate it less into their revision.

3. Does Mandating Affect Students' Emotional Engagement With Écree?

As seen in section 1, the EO class was likely to evaluate Écree feedback more positively than the EM class did. This difference between EO and EM classes in their evaluations of Écree may indicate that these two groups have different emotional engagement with Écree. The analysis of the interviews data with the focal students reveals the EO and EM classes have different views on the usefulness of Écree and satisfaction with Écree.

1) Usefulness of Écree

The four participants in the EO class mentioned the usefulness of Écree more often than their counterparts in the EM class, whereas students from the EM class referred to Écree as "not being useful" more often than not. The EO group referred to Écree's systematicity, speed, and its role as a surrogate reader as its key advantages.

> Excerpt 1: Semin from EO
> When human raters give feedback, they do not apply the same guidelines to each paragraph. They find some mistakes in this paragraph and other mistakes in that paragraph, so they may miss some others. But Écree applies a fixed framework to all the paragraphs and checks everything.... Écree applies the same criteria all the times, so its feedback is accurate. Moreover, the feedback is so immediate, it does not take more than three minutes.

Semin compares Écree feedback with humans and highlights its positive effect citing the systematic application of

the guidelines through the whole essay as well as the immediacy of providing feedback. On the other hand, both Minho and Jinhee state that Écree serves as a surrogate reader who substitutes for a human reader as follows:

> Except 2: Minho from EO
> Frankly speaking, I hope someone will refine my writing before submitting it, but in reality, I don't have someone available to read it all the time. But Écree provides feedback immediately, in less than 10 minutes.

> Excerpt 3: Jinhee from EO
> I rather consider [Écree] as one more opinion. I don't accept Écree feedback too seriously nor too importantly. I just use Écree to see if my writing is grammatically correct or I think Écree as another pair of eyes….sometimes I have to cut out something between sentences. Then I'm not sure whether the flow is okay. From my point of view, the flow looks okay, but from others' perspective, it may not. But I don't have many friends I could ask to read my writing, so I enter my writing into Écree, which checks the flow.

Unlike these favorable attitudes towards Écree feedback's usefulness, the mandatory group expressed discontentment. They voiced concern about its inaccuracy, its inability to consider the writer's intention, and inconsistency.

> Excerpt 4: Yejin from EM
> I thought that I wrote a thesis statement and restated it in the conclusion, but Écree said I did not. If my writing gets a little long, Écree says too much detail or if writing gets a little short, Écree says not enough contents this time…. One time I accidently hit the enter button. Then Écree considered that as a body paragraph and pointed out problems wrongfully. So, I thought that the machine is a machine.

As Yejin points out Écree's limits to understanding the writer's intention, Wonjoon also pinpoints the impossibility of negotiating with Écree as one of its greatest disadvantages:

> Excerpt 5: Wonjoon from EM
> I definitely put a topic sentence, but Écree says there is no topic sentence by putting the part in red. I can't tell Écree this is my topic sentence because it is a machine. So, I tried revision a couple of times, but Écree keeps the part in red. So, I gave up.

These negative views on Écree feedback seem to lead to their belief that its feedback is not trustworthy as Youngmee's experience led her to believe.

> Excerpt 6: Youngmee from EM
> I just made minimal changes such as typos and reentered my writing into the Écree program. Because it was my first time, I put it after three minutes. Several feedback came after each paragraph, but they were quite different from before. Before Écree said something is good, but later it said that is not good, although I didn't change the content at all. Écree is not reliable.

2) Satisfaction with Écree

These two groups' differing attitudes towards the usefulness of Écree could have contributed to difference in their level of satisfaction with its feedback as well. The participants in the EM group mentioned their dissatisfaction with Écree feedback more frequently than their counterparts in the EO group. In only one occasion, dissatisfaction with Écree feedback was noticed in EO, but fourteen complaints about dissatisfaction with Écree were made by the four interviewees from the EM class.

The course instructor related EM group's dissatisfaction with Écree to a lack of specificity of the feedback and mismatch between Écree and the course objectives. In Excerpt 5, for example, Wonjoon complained about Écree's feedback on his topic sentence. One topic sentence that Écree rated as "weak" was "*There are some classes that are both theoretical and practical*" (Wonjoon, Essay 2, draft 2). When scoring the final draft, the instructor noted that Wonjoon should take a stance, such as, "*The best classes are both theoretical and practical*" (Wonjoon, Essay 2, final draft, feedback). This example illustrates that although a problem area had been identified by Écree, the specifics of how to fix it were not clear to the students, and this led to frustration. On the other hand, the EO students—because they had not been required to use Écree—could simply ignore Écree's feedback, or treat it as another set of eyes (see Excerpts 2 and 3).

Additionally, the instructor assumed that mismatches between Écree and the course objectives were more likely to cause frustration to the EM class than to the EO class. He cited one issue related to the thesis statement of essays. The course objectives stated that a thesis statement should include a three-part writing plan that outlines the body paragraphs. However, Écree did not assess the thesis statement along these lines. For example, in her first essay, Youngmee's thesis read, "*The pronunciation of Spanish is really attractive and it creates a desire to learn more*" (Youngmee Essay 1, draft 1). Écree rated this as "clear" (green). However, Youngmee's thesis did not have a clear writing plan as taught in the course which cost her points on her essay's score. As seen in Youngmee's case, the instructor interpreted that the EM group uploaded more often, more uploads mean more feedback. Because sometimes the feedback did not match the instructor's feedback or the class contents, however, the EM group grew frustrated whereas the EO students could ignore either the feedback or simply not use Écree (e.g., Sumie).

The frustration with Écree feedback led to one further observation: EM students said that Écree should not be mandatory, while EO students said it would be a good idea to make students use it. All four EM students interviewed said it would not be a good idea to mandate Écree. Mina and Youngmee suggesting have it as an available resource, while Wonjoon and Yejin objected totally. The EO students said it would not be a bad idea because students could verify their essays before submitting them. This finding is in line with one of the points raised by Jones, Richards, Y. Cho and Y. J. Lee (2019) who in their survey of students' perceptions on English learning in the Fourth Industrial Revolution (4IR)

found that students who were more familiar with technology tended to think less positively of it. Jones et al. (2019) suggested that the technology-familiar students may have tried using it to learn English and not achieved as much as they thought, whereas those less familiar with technology were imagining its capabilities (p. 62). This may be the case with Écree as well.

V. DISCUSSION AND CONCLUSION

This study compares and contrasts two groups: one required to use an AWE system called Écree, and one for whom Écree use was optional, in terms of their overall engagement, in particular, focusing on their behavioral and emotional aspects. The number of uploads shows that EM grouped used Écree more frequently than EO group, but the survey data reveals quite a different result, that is, EO group assessed the extent to which they incorporated Écree feedback into their revisions more highly than EM group. These confounding results between their actual behaviors and their perceptions on Écree, are confirmed and also accounted for by the analysis of the focal students' behavioral and emotional engagement with Écree. The analysis of their written drafts reveals that the EM group used Écree more frequently and incorporated its feedback into their revision more than the EO group; however, towards the end of the semester, their usage of Écree as well as the extent to which its feedback was reflected into student revision decreased. The analysis of their interview data indicates that this decreased use of Écree may relate to their perceptions of it. While the EO group's perceptions were quite positive about usefulness of Écree, the EM class had doubts about its usefulness. The EO group tended to acknowledge Écree's usefulness—that is, Écree feedback was not only systematic and fast, but also it substitutes for a human reader so as to provide another reader's perspective on their drafts. The EM group, however, tended to have negative views on Écree, often emphasizing its disadvantages such as its inaccuracy, inconsistency, and lack of understanding of the writer's intention. The mandatory group's negative views on AWE further lead to their lack of motivation (as in the case of Yejin, Excerpt 4) and demotivation (as in Wonjoon, Excerpt 5) to use Écree in the future. The participants' quite distinct views on the use of AWE depending on whether they were required to use it or not, as suggested by this study, requires further investigation.

This study confirms what other studies have found about the effectiveness of AWE: its effectiveness depends on how it is integrated into a writing class (CCCC Executive, 2006; Chen & Cheng, 2008; Li et al., 2015). Chen and Cheng (2008) implicate that the mode of use of AWE matters in examining its effectiveness in a writing class, and thus, if it is not integrated into a writing class in an appropriate way, as CCCC and Cheville (2004) argue, the student may develop negative views towards the use of AWE.

One pedagogical implication of this result relates to the use of feedback. In many cases, a classroom instructor would mark the error to tell the student what to fix (indirect correction), or they may even make the corrections for the student to recopy (direct correction). If the student did not understand, they could simply raise their hand and ask the teacher. In this respect, the feedback is highly specific. In the case of Écree, students cannot ask. Wonjoon's Excerpt 4 (Section III.1) and weak thesis statement (Section III.2) show how this played out for him. He wanted to *tell* the machine that he had a thesis statement, and he could not figure out what was meant by Écree's "weak" thesis statement feedback. The implication is that students need to be trained in how to interpret the clues given by the AWE in a way that improves their writing. This reflects J. Park's (2019) finding of missed and incorrect grammatical feedback from *Grammarly.* Écree also missed some important points and gave incorrect feedback on others Neither AWE is infallible. Teachers, therefore, need to familiarize themselves fully with what an AWE can and cannot do, and use previous students' sample essays to compare the AWE with their own feedback. This way they can train students to identify which feedback they should take on board and which they should leave behind.

Another item to note here was that during peer review, due to the design of the study, EM students had a chance to discuss their Écree feedback with each other, but according to the classroom recordings, they rarely did. For the most part, reviewers simply gave their own feedback to the writer and ignored what Écree had said. The only time EM students mentioned Écree was to complain – often to the instructor. EO students did not have the opportunity to use Écree before peer review, but could, if they used Écree, compare their peer feedback to Écree's after uploading. There is a potential confounding factor; if the peer feedback were poor, Écree's feedback might seem better. However, no student in either EO or EM reported Écree giving clearly superior feedback than their peers did (There were no answers of 5 or 6 on survey question 19a). Additionally, if (in the EO condition) peer feedback were excellent and the students integrated it into their draft before checking with Écree, Écree's feedback may well be more on point, and therefore seem better. This bears future investigation. Even if future research finds this to be the case, it still makes the argument that AWE would be best be used in conjunction with human feedback—a position similar to Chen and Chang (2008) and in line with our call for both teacher and learner training before using.

Another related issue is familiarity with writing and with technology. In this study, those who had been mandated to use Écree both used it more and liked it less. It is possible that familiarity bred contempt. Jones et al. (2019) indicated that those who had used more technology to learn English also felt it was less helpful, and surmised that students had tried to use English language learning technology and were less successful than they had hoped, therefore downgrading its potential. Our findings are in line with this.

However, Jones et al. (2019) also note that lack of knowledge of technology was among the biggest obstacles to its use. This leaves the teacher in a difficult place. They must first familiarize themselves with the technology, and with how their students might perceive it. Then develop teaching strategies and materials to help students make better use

of technological resources, including AWE. In the case of Écree, we suggest the teacher give content related feedback on a number of sample essays then upload them to Écree and compare the two types of feedback given. Analyzing the similarities and differences will give the teacher an idea of what students should look for and what they should ignore.

Because this study was based in two pre-existing classrooms, and the number of focal participants is small, the findings cannot be over-generalized. They do align with other work and point to the need for more study on the human-AWE interface. First, we argue for a policy that the instructors and students get experience with the AWE before "going live". We also recommend that Écree itself produce materials for teachers along these lines—either video, sample feedback worksheets, etc. that will increase the students' capacity to understand the implication of the feedback. Zhang's (2017) study of one participant showed how she engaged with the feedback she received, but she was in all cases engaging with the score. With Écree, the student does not see their score, and engages with colored words instead. Because these words are not as precise as what their instructor would give them, trying to react to them, without training, led to frustration. In the mandatory group, which used Écree more, this frustration led to negative feelings about Écree, whereas EO group members could avoid Écree if frustrated.

Second, as the participants in the mandatory class pointed out, the impossibility of negotiating with Écree over its feedback, one of the greatest obstacles to the wide use of AWE may be that current AWE does not have human readers' capacity "to perceive what works and to imagine what might work better" (Cheville, 2004, p. 51). AWE has been improved in a drastic way over the past few decades, but it should have a long way to go so as to understand the writer's intention as a human reader will do. Until that development has been made, this study implies that AWE programs can be used as one of the writing resources available to students in a writing class as suggested by Shermis and Burstein (2003), and Warschauer and Ware (2006), with the caveat that teachers and learners understand what type of writing gets what type of feedback.

REFERENCES

Armel, D., & Shrock, S. A. (1996). The effects of required and optional computer-based note taking on achievement and instruction completing time. *Journal of Educational Computing Research, 14*(4), 329-344.

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

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater v.2. *Journal of Technology, Learning, and Assessment, 4*(3), 1-30.

Baron, D. (1998, November 20). When professors get A's and the machines get F's. *Chronicle of Higher Education*, A56.

Brandl, K. (2012). Effects of required and optional exchange tasks in online language learning environments. *ReCALL, 24*(1), 85-107.

Burstein, J., Chodorow, M., & Leacock, C. (2004). CriterionSM: Online essay evaluation: An application for automated evaluation of student essays. In J. Riedl & R. Hill (Eds.), *Proceedings of the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence,* Acapulco, Mexico (pp. 3-10). Menlo Park, CA: AAAI Press.

CCCC Executive Committee. (2004). *CCCC position statement on teaching, learning, and assessing writing in digital environments.* Retrieved from https://cccc.ncte.org/cccc/resources/positions/digitalenvironments

CCCC Executive Committee. (2006). *Writing assessment: A position statement.* Retrieved from http://www.ncte.org/cccc/resources/positions/writingassessment

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.

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

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

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

Farrell, T. S. C., & Kennedy, B. (2019). Reflective practice framework for TESOL teachers: One teacher's reflective journey. *Reflective Practice, 20*(1), 1-12.

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). Automated essay scoring: Applications to educational technology [Electronic Version]. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning.* Retrieved from http://imej.wfu.edu/articles/1999/2/04/index.asp

Garland, K., & Noyes, J. (2004). The effects of mandatory and optional use on students' ratings of a computer-based learning package. *British Journal of Educational Technology, 35*(3), 263-273.

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), 4-44.

Heit, J., & Donaldson, R. (2018). Écree. Retrieved from https://www.ecree.com

Jones, S., Richards, A., Cho, Youngsang, & Lee, Yoo-Jean. (2019). Digital technology in the 4IR and the future of English learning from the perspective of

Korean EFL university students. *Modern English Education, 20*(1), 53-70.

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.

Leki, I. (2006). "You cannot ignore": L2 graduate students' response to discipline-based written feedback. In K. Hyland & F. Hyland (Eds.), *Feedback in second language writing: Contexts and issues* (pp. 266-286). Cambridge: Cambridge University Press.

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), 1-18.

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

Park, Junhee. (2019). An AI-based English grammar checker vs. human raters in evaluating EFL learners' writing. *Multimedia Assisted Language Learning, 22*(1), 112-131.

Papageorgiou, S., Tannenbaum, R. J., Bridgeman, B., & Cho, Y. (2015). *The association between TOEFL iBT® test scores and the Common European Framework of Reference* (CEFR) levels (Research Memorandum No. RM-15-06). Princeton, NJ: Educational Testing Service.

Shermis, M. D., & Burstein, J. (Eds.). (2003). *Automated essay scoring: A cross-disciplinary perspective.* Mahwah, NJ: Lawrence Erlbaum Associates.

Shermis, M. D., Burstein, J. C., & Bliss, L. (2004, April). *The impact of automated essay scoring on high stakes writing assessments.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Shermis, M. D., Garvan, C. W., & Diao, Y. (2008, March). *The impact of automated essay scoring on writing outcomes.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York.

Vantage Learning. (2007). *MY Access!® efficacy report.* Retrieved from http://www.vantagelearning.com/learning-center/research/#MyAccess

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: An International Journal, 3*(1), 22-36.

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

Williamson, M. M. (2004). Validity of automated scoring: Prologue for a continuing discussion of machine scoring student writing. *Journal of Writing Assessment, 1*(2), 85-104.

Zhang, Z. (2017). Student engagement with computer-generated feedback: A case study. *ELT Journal, 71*(3), 317-328.

APPENDIX 1

Survey Questions

Part 2: Écree

	Question	1	2	3	4	5	6
13	(D or A) I revised the parts of my essay which écree did not classify as green. Reason: _____						
14	(D or A) I was able to fully apply écree's feedback to revisions of my essay. Reason: _____						
15	(D or A) I trusted écree's feedback about: (a) Language (Grammar and spelling) (b) Organization (Thesis statements, etc.) (c) Content (Examples, details, ideas) Reason: _____						
16	(D or A) Écree feedback helped me give better feedback during the PRC. Reason: _____						
17	(D or A) Écree should give me a score as well as red, yellow, and green parts. Reason: _____						
18	How often did you disagree with écree feedback? Reason: _____						
19	(D or A) Écree feedback was more helpful than: (a) my peer reviewer's feedback. (b) the professor's feedback (c) my own analysis Reason: _____						
20	(D or A) The feedback from écree was sometimes not relevant to my point. Reason: _____						

21. In what way(s) was écree useful for your development as a writer?

_____
_____
_____
_____
_____
_____

22. Would you recommend that the professor use écree in next semester's class? Why or why not?

_____
_____
_____
_____
_____
_____

APPENDIX 2

Interview Questions

1. Did you use Écree? How many times? If you didn't use Écree, why didn't you use it?
2. How did you feel when you read the feedback? Explain.
3. Were you able to understand and use the feedback that Écree gave you? Was Écree feedback helpful for you? Why or why not?
4. Would it be better if you could see the score the machine gave you? Why or why not?
5. (Mandatory class only) Should the professor require you to use Écree? (Or is it better to be optional?)
6. (Optional class only) Did you talk about the Écree feedback in your peer review circle? Why or why not?
7. Is there anything else you'd like to tell me, or the professor about the peer review circle, Écree, or the class in general?
8. Would you like to use Écree again next semester (or the next time you take a writing class)?