The Effect of Using Automated Essay Evaluation on ESL Undergraduate Students' Writing Skill

Ebtisam S. Aluthman¹

¹ Department of Applied Linguistics, Princess Norah bint Abdulruhman University, Riyadh, Saudi Arabia

Correspondence: Ebtisam S. Aluthman, Department of Applied Linguistics, College of Languages, Princess Norah bint Abdulruhman University, Riyadh, Saudi Arabia. E-mail: esaluthman@pnu.edu.sa

Received: July 24, 2016Accepted: August 14, 2016Online Published: September 23, 2016doi:10.5539/ijel.v6n5p54URL: http://dx.doi.org/10.5539/ijel.v6n5p54

Abstract

Advances in Natural Language Processing (NLP) have yielded significant advances in the language assessment field. The Automated Essay Evaluation (AEE) mechanism relies on basic research in computational linguistics focusing on transforming human language into algorithmic forms. The Criterion® system is an instance of AEE software providing both formative feedback and an automated holistic score. This paper aims to investigate the impact of this newly-developed AEE software in a current ESL setting by measuring the effectiveness of the Criterion® system in improving ESL undergraduate students' writing performance. Data was collected from sixty-one ESL undergraduate students in an academic writing course in the English Language department at Princess Norah bint Abdulruhman University PNU. The researcher employed a repeated measure design study to test the potential effects of the formative feedback and automated holistic score on overall writing proficiency across time. Results indicated that the Criterion® system had a positive effect on the students' cores on their writing tasks. However, results also suggested that students' mechanics in writing significantly improved, while grammar, usage and style showed only moderate improvement. These findings are discussed in relation to AEE literature. The paper concludes by discussing the implications of implementing AEE software in educational contexts.

Keywords: Automated Essay Evaluation, academic writing, language Assessment, Saudi ESL undergraduate students

1. Introduction

Scholarly research has long depicted academic writing as a complex socio-cognitive construct involving a continuum of activities (Omaggio, 1993). Many national and international education standards have placed more emphasis on the Common Core State Standards (CCSS) (OECD, 2012; Ananiadou & Claro, 2009). The CCSS necessitates that students manifest highly proficient writing skills, including summary and synthesis. The CCSS underlies the mechanisms of most, if not all, the available standardized English proficiency texts, such as ITELS and TOEFL.

The issues involved in both academic writing improvements and assessment are often just as complex as the construct of writing itself. Scholarly research into writing shows that instructional feedback is the most efficient way of improving writing (e.g., Wilson et al., 2014; Graham et al., 2011; Anderson et al., 1985). Hattie & Timperey (2007) define instructional feedback as providing information that indicates levels of correctness as well as a means of improvement. More pertinent to the aims of the present paper is the role of individualized and instant feedback given to students that is indicated in the scholarly research as improving the students' writing proficiency (e.g., Covill, 1997; Etchison, 1989; Fitzgerald, 1987). Unfortunately, this process places an enormous workload on classroom instructors who are in charge of reading and correcting a large number of essays per any writing assignment. As a result, teachers may be unable to assess and correct students' written work as often as they wish. Meanwhile, instructional feedback related to writing performance is not only time-consuming but also problematic due to inconsistency and instructor-centeredness (Wilson et al., 2014; Grimes & Warschauer, 2008; Lee et al., 2009). The degree to which instructors provide accurate and comprehensive feedback also remains unclear. As Zamel (1985) reported, instructors are often inconsistent, arbitrary and sometimes contradictory.

In response to scholars' research into the effects of instructional feedback on writing proficiency, recent studies

have placed great emphasis on utilizing NLP and Artificial Intelligence (AI) in providing students with both an automated holistic score and immediate feedback. Scholar research in the AEE area began in 1960s and has been growing rapidly (e.g., Burstein et al., 1998; Burstein et al., 2004; Shermis & Burstein, 2013). The rationale for using these computing methods is to provide a valid, consistent and time-saving system for assessing student errors and for providing instructor feedback and comments. A number of AEE systems have sought to provide both automated individualized scores and feedback. Research into AEE varies in scope and includes studies that have explored the usefulness of AEE techniques on writing proficiency (e.g., Rudner & Liang, 2002; Attali, 2004; Franzke et al., 2005; Wang, 2011). And those that have compared human to automated evaluation in terms of reliability and validity (e.g., Cohen et al., 2003; Kulik, 2003; Wang & Brown, 2007; Bejar, 2011; Bejar, 2012).

As some educational institutions are implementing various AEE systems, further research is necessary to investigate the effectiveness of these recently developed systems in terms of improving students' overall writing. Within the field of AEE research, little attention has been given to the implementation of AEE systems in the English as a Second Language (ESL) context. Therefore, the present case study examines the effect of utilizing AEE on Saudi undergraduate ESL writing performance, an area that has never been investigated in this body of literature. This research is guided by the hypothesis that AEE tools are beneficial in terms of overall ESL writing improvement. Specifically, two main issues are addressed within the scope of this investigation: a multi-level quantitative analysis investigates the positive effect of utilizing AEE on ESL undergraduate students' writing proficiency as well as the kinds of errors that more dissipate through the use of AEE. The following review of the literature provides a brief overview of AEE advances, including its history, different systems and potential applications as well as empirical studies conducted in this area.

2. Literature Review

2.1 Automated Essay Evaluation

Considering the complexity of writing assessment and improvement, researchers in the field of AEE have worked on developing a variety of computer-based systems that automate the process of both scoring and providing feedback. Practical AEE efforts date back to the early 1960s when researchers began to seek to develop automatic scoring applications. Page's article, "The Imminence of Grading Essays by Computer" (cited in Shermis & Burstein, 2013), began the tradition of AEE research. A stable working version of Page was released in 1973 (e.g., Shermis & Burstein, 2013). According to Page, technology can be used as a tool by instructors burdened with hours of grading writing assignments. The concept may have been before its time: word processing packages were not to become available until the beginning of the next decade, leading to many objections regarding the idea of displacing human raters.

Pioneering investigations into the area of AEE were initiated in the 1980s with the work of the Writer's Workbench (MacDonald et al., 1982). Based on the reviews of Warschauer & Ware (2006) as well as Ebyary & Windeatt (2010), Table 1 provides a detailed description of the most well-known AEE systems in terms of producing companies, software engines, targeted areas of writing assessment, statistical approaches to evaluation and types of scoring and feedback

Company	Software	Areas to Be Measured	Statistical Approaches	Scoring	Feedback
MI Measurement Incorporated	PEG	Fluency, diction, syntactic complexity	Regression	Holistic and trait scores (Note 1)	Feedback
ETS English Testing Centre	The e-rater®	Grammar, usage, mechanics, style and organization	Regression	Holistic score	Detailed Individualized feedback
Pearson Knowledge Analysis Technologies	IEA	Content, mechanics and style	Latent semantic analysis regression	Holistic and trait scores	Limited individualized feedback
Vantage Learning	Intelli-Metric —tm	Cohesion, coherence, content, discourse, syntactic complexity, variety and accuracy	Artificial intelligence	Holistic and trait scores	Individual feedback

Table 1. The most well-known AEE systems

The AEE system chosen for the present study is e-rater®, the AEE platform manufactured by the Educational testing Service (ETS) that provides formative feedback along with an automated holistic score. This system's automated scoring application has been used for the TOEFL IBT as an independent rater for the purpose of scoring the writing independent writing task (see Enright & Quinlan, 2008 for the use of the e-rater® system in the TOEFL IBT independent writing task). The following section summarizes empirical studies into the effects of AEE on writing improvement.

2.2 Automated Essay Evaluation and Writing Improvement

Significant interest has grown gradually but steadily over the last decade in the field of AEE research, particularly in terms of investigating the effect of utilizing AEE systems on improving writing. Underpinning this interest is many studies indicating the critical role of formative feedback on students' writing improvement (e.g., Black & William, 1998; Vygostsky, 1978). This interest grew sharper when scholarly research documented that the agreement between human raters and the e-raters® evaluation system is 87%-97% (Burstein et al., 2003; Valenti et al., 2003). However, most of the empirical research conducted to investigate the effect of utilizing AEE systems relates to English as the native language rather than English as a second language. Studies exploring the effects of utilizing AEE systems on writing have varied in their contexts, methodological design and participants.

Many studies documented in the AEE area of research have investigated the effect of AEE systems on grade 6-10 students within the L1 context. Shermis et al. (2008) utilized a hierarchical linear study design to examine the effect of the e-rater® system on writing improvement in this group of students. Their analysis includes different measures, holistic scores, word counts, word usage and errors committed in grammar, style, mechanics and usage. The data included the final drafts of seven different student essays over 11months. Results indicate a significant improvement in the students' writing. Recently, Wilson et al. (2014) examined the effect of immediate instructional feedback provided by the PEG system on the overall writing quality by investigating data from grade 4-8 students. The researchers applied a three-level hierarchical linear study to identify the effects across different revisions. They found that certain groups of students scored higher on their first drafts, including females and more proficient writing students. They also observed that there was no significant transfer effect in subsequent prompts.

Applying an experimental study, Kellogg, Whitford, & Quinlan (2010) examined the influence of individualized automated feedback on writing quality using the e-rater® system. They randomly assigned undergraduate students to three different groups depending on the amount of feedback they received from thee-rater® (zero feedback, moderate feedback and constant feedback). Students who received constant feedback demonstrated reduced errors in grammar, mechanics, usage and style in their final drafts. Ebyary & Windeatt (2010) applied questionnaires, interviews, and focus groups to investigate the effects of using the Criterion® systems with 549 Egyptian trainees and EFL instructors. The investigation focused on attitudes towards using the Criterion® system and whether there was a noticeable influence on the writing strategies used by students. They noted the existence of generally positive effects on the students' planning strategies in their revised drafts as well as positive attitudes towards using AEE systems.

Few studies have investigated which areas of automated feedback are most effective in enhancing students' overall writing (grammar, lexis or organisation/structure).Focusing on aspects of content and organization, Lee et al. (2009) applied a web-based evaluation application that provides students with instant individual feedback on both content and organization. Through an experimental study with twenty-seven students assigned randomly to experimental or control groups, Lee et al. (2009) reported that there were no significant effects of applying a web-based critiquing tool with adult EFL learners in regard to both the content and the organization of students' writing.

In summary, the relevant empirical studies in the field of AEE systems suggest positive effects of using AEE systems on overall writing improvement. However, only a handful of studies have explored the effectiveness of AEE systems on EFL students' overall writing quality. Therefore, further research is required to test the efficiency of these systems and to determine which areas in the writing construct can be effectively improved via AEE systems.

In ESL writing research, as Silva (1993) argues, the task of writing in L2 is generally more constrained and difficult than L1. Pertinent to the present study, Process Approach has dominated the field of L2 writing since the 1970s. It focuses on the processes of idea generation, drafting, giving and receiving feedback, and revising. Related to L2 writing improvement, Ferris (2011) emphasizes the positive role of corrective feedback in L2 writing. According to Ferris (2011), the comprehensive marking of errors and strategies for correcting them are preferable to direct correction. However, the degree to which L2 instructors provide accurate and comprehensive

feedback is unclear. Zamel (1985) reported that instructors are often inconsistent, arbitrary and contradictory. It is argued that the AEE application used in the present study has a significant influence on the standardization of the process of both evaluating writing errors and providing feedback to students.

Though AEE is increasingly adopted, it has never been used in any Saudi Arabia educational settings. This study presents an account of an empirical study conducted at Princess Norah bint Abdulrahman University where an AEE system was used for the first time. Investigating the effectiveness and implications of utilizing AEE in the current setting will contribute to our understanding of AEE as well as pedagogical enhancement of writing courses and curriculum design. For practicality and ease of use, the Criterion® system was chosen for study administration and data collection. In the next section, an overview is provided of the Criterion® scoring mechanism.

2.3 The Criterion® System

The Criterion® system combines an automated scoring feature and corrective feedback. Two complementary functions have been developed through NLP methods: critique and e-rater® version 2.0. Critique consists of program software designed to provide immediate evaluation and feedback in terms of grammar, mechanics, usage, discourse and stylistic features. Meanwhile, the e-rater® implements statistical measurements in extracting linguistically-based features from an essay based on corpora that has been processed and inputted into the application. The Criterion® application provides a holistic score on the writing assignment by applying a statistical model determining the association between extracted features and overall writing proficiency. The instant score provided by the Criterion® system is based on four categories of features. As illustrated by Criterion Online Writing Evaluation Services, these core categories include grammar, mechanics, usage, and style.

The Software Sustainability Institute has undertaken a software evaluation of Criterion®, releasing their detailed multi-criteria report in November 2011 regarding sustainability, maintainability and usability of the system. This evaluation report provides a high-level description of the audience for the software and its inner workings. It also gives a high-level overview of the software that consists of clear, step-by-step instructions. Criterion® is a web-based instructor-led instrument that provides learners with a variety of tools to plan, compose and amend their writing assignments. It offers a virtual classroom where instructors can create classes, design assignments, make announcements and give diagnostic feedback. According to the Criterion® developers, the program has been used in a variety of educational settings including primary schools, high schools as well as higher education institutions. In the current study, an AEE system was integrated into an academic writing course for the first time. It is hoped that the remarks and observations obtained from this study will contribute to further investigation in the AEE field. The following sections outline the use of the Criterion® in the present research.

2.3.1 How the Criterion® System Works

The Criterion® system offers a virtual classroom where instructors can create classes, design assignments, make announcements and give diagnostic feedback. In a friendly easy-to-use, The Criterion® homepage offers a variety features as illustrated in the following figure.

Criterion	HOME CLEAT SUBVCES RESOURCES HELP Janis Strohl Instructor Sign Out
Welcome acts and instructions to user from the Citarion system. This message may be edited by the ECA at any time and speaks individually to each Connect to an additional methodor, a higher role in your methodor, or a new cleas as an instructor or addent Notices As you create classes and add asset to records to dealer.	nde of Student, Instructor, or Clevel Admin. Signments, they will be listed
Classes Add instructions how to classes table for instructors below Image: Classes Image: Cla	From the row of icons you can perform the following tasks: View – Check the box in front of the name of the class, and view current activity within the class. Add – Create a class. Criterion automatically generates the access codes for instructors and students.
Class Grade Instructors A Demo Client DFriedman School	Edit – Edit class information. Announcements – Generate a dated and/or default
test class Grade 11 Der Studiet1 Mr. Teil münucuto2 Jacket Jacket Australingender Instructure AAST Mis. Jans Bruhi	announcement to be sent to all class members. Connect – Search for an additional instructor from your school by name and connect him/her to your class. Disconnect – Disconnect an instructor currently registered to your class.
Legal / Privary & Senate / Alast Criterion Complete 2213 (r. 19. 17), An right senarrow The Senator of Constitution of Constitution and Chill Children and Antonia and Children and a difference and Children and a senator and the senator and the senator of Children and Senator and Se	Libroing Londong'

Figure 1. The criterion® homepage

Using the basic functions of the Criterion® is conducted through five main phases.

1) Preparing writing topics

The Criterion® system offers a variety of topics, levels and modes (persuasive, informative, expository, narrative, and argumentative). Instructors can also choose other topics and design their own prompts. Time limit and number of attempts are also chosen in this phase. The following figure illuminates these different options.

Criterion	HOME ; CLENT SERVICES ; RESOURCES ; HELP Jamis Strohl Instructor Sign Out
Strohl Sample Testing Announcement	Class: Strohl Sample, A Demo Clerit, DFredman School • Go
Activity Assignments Roster Reports Add / Edit Assignment Add instructions for how to build an assignment. Image: Comparison of the state of the st	 Choose whether or not to select an administrator assignment. Choose an assignment type from the drop down menu: Topic Library, CCSS (Common Core State
Select an Administrator Assignment :	 Standards, Scored Instructor Topic or Text Editor. Choose Level, Mode, and Prompt from the dropdown menus. Click View Topic Library to view full text of library prompts. If you choose a prompt from the Topic Library dropdown, the name and text of the prompt will appear. Add up to 7 URL reference links to any assignment for multi-stimulus prompts, custom plan templates, etc.
Reference Linit: Display Name: Add	

Figure 2. Creating a writing assignment

2) Composing the writing assignment

Students log in and plan, compose and revise their writing assignments. Different planning templates are available to help students in prewriting strategies.

	Creating a Plan	
@ Criterion	e Rectangular Spin	neman i Anna i Arte
et electrica de la properties de la prop	To begin an assignment, the stuc clicks on the Plan tab. • By default, the Outline plan i open unless you have specifi plan. • A student can choose a diffe plan, by clicking on any of th other hyperlinks. • Students type directly into th plan.	dent

Figure 3. The plan/response window

3) Submitting the assignment and receiving an immediate score

Once students submit the writing assignments, an automated score and diagnostic feedback are given within 10 seconds.

$k \neq 10$, will λ large the local state. Also also accord over they also the wave, where $k \neq 10$, where $k \neq 10$, and $k \neq 10$. Since the state $k \neq 10$, and $k \neq 10$, and $k \neq 10$. Since the state $k \neq 10$, and $k \neq 10$.
Another the second se
Click to view
the full scoring guide
envertices Organization, Development and Style
Level 5 - Advanced

Figure 4. Viewing feedback

4) Revising and re-submitting the assignment

Specific feedback of each category of errors is given.

Licing	Eaghack Dran	Down Monus
Using	reeuback Drop	Down Menus
@ Criterion		THE I LANDREN I MOUNT MAIL
batches	A Rectangular Snip	then test blas, i Dene Tant, Minather Schart - 1
Dense Talet Note: Serve Tale		reeks old. They are extremely cute but can
Asseguration		You have used there in this
increasing in carrier and store equal and	anonyclass Tanta Behavior Athengo 8.1	f you sentence. You may need to use o ital whods they're instead.
Manual Institution		there for you and if there not than you can
	Any trait category followed by a	
	number in parentheses,	
Seguration & Realization (1) Barge (8) Michaelas	indicates feedback is present.	Click on the category, and
The day that I got my lotner was a day that I to	J.	the error is highlighted in
grow much require and freedom from my party arrival and drop's come or said's array from of these	or green no emport from other adult detroits, her more they taken the mesone of engeneticity that com-	roll-over note is provided.
Contrain constructions for forgening detroins. The n two have loaned we will pace on to the fortune g. Proposition form ward.	to paug not and having a grow time. Using 10 to the failur bulles we progress into the singer of add here children. "The oil clube "Do arms others and they will do arms you" applies to the fluctuating one	Hand, Vi a Intellige for
Everyone the age of [2] and older was a transage	Name your 2015 more likely that is high school they partial fessely. This is far case of any same Onto, rappe above that above angles. We way to ball that are that fiction is out of the bars when sho way to	viso will have 30 years and data Decouples. 'In high,

Figure 5. Specific trait feedback

Based on the provided diagnostic feedback, students revise and fine-tune their writing. Reviewing errors involves identifying the category error, numbers of errors in each category, highlighting errors and providing suggestions and advices. All these procedures are illustrated in the following figure.

Plan Response Re	esuits	
Leadership		
A leader is defined as a person who d have which would encourage people to	irecta, commands or guides a group follow them. Be as specific as oceas	or activity. Think about someone you know who has proven to be a great leader, either in leading a country or team or organization. Write an essay about this leader, explaining the characteristics the bits in illustration the attrobules that this leader has exhibited which would lead others to respect admise and follow them.
		ana na mana ang na anatana na na anatana na anatana ang nana ang nana na ang ang ang anata ang nana nana
		Date
Organization & Development	Grammar (1) Usage (3)	Feedback category
8 0 A 4 5	Fragment or Missing Comm	dran dayun
	Run-on Sentences	aropaown.
iome people have a charisma	Garbled Sentences	ourages others to follow them. They display confidence, compassion, and can work cooperatively with others. In its brief history our country has had a
sumber of great leaders. Am	Subject-Verb Agreement	out as a man of integrity, a man of action, a man of the people.
One of Lincoln's important pe	III-formed Verbs	ten actively to other people. Not only did he seek the ideas of his fellow politicians, but he made time to listen to the common people both in person and
hrough thoughtfully reading a	Pronoun Errors	tory books are filled with letters he received during his time in office and with his thoughtful responses. He was known as a quiet man who would listen with
nterest to what others had to	Possessive Errors (1)	In making decision
The ability to lead by example	Witness or Hissian Wand	ting leadership tra to show possession. Ex our to cause in KOIIOVER NOTE. in his power to spare his son from the war.
vas a savwy politician who un pole time to talk with him, pla	wrong or Missing viora	be necessary to achieve the nations acals and to bind the nation's woonas. Luncous is remembered as a name figure, who pulled his small son to his lap, an
He withstood great sadness,	Proofread This!	the store a parent s greet in the sess of a bold win great usging. Sharing the common experiences of the masses made others neet that he was sone of them. strength, and dignity.
the state of the s		
incom was a man or action. A Proclamation, knowing that it i	is president during the Civil v might lead to the secession of	Tar, he was presented with extremely duthe the South. As commander-in-chief he was the local state and most
ostly war in our history. Whe	en the war ended, he took acti	on to bind the nation's wounds by allowing Southern soldiers to return to their homes and proposing plans to rebuild the country. He wore the yoke of
esponsibility for the fate of ou	ir country.	
	Click hore	and in revising an assau

Figure 6. Reviewing errors

5) Personalized feedback is added.

In this phase, instructors add their comments on students' submitted assignments, create dialogue and give suggestions.

Note that from this page, you can view the student plan, response and results.			-		Dialogue
		Acceptable Teen Behavior	,	Click the dialogue but the right hand side.	ton on
Plan Response Results			_		Export
Acceptable.Teen.Behavior In an effort to set guidelines for acceptable trenage behavior, p Hou much freedom and responsibility should trenagers have in Additional Material:	arents and other adults in authority often making their own decisions? Esplain your	Enter your feedback for the student and click Send. The student will be able to view the feedback by clicking the dialogue button from his Response screen. A scrolling list of all dialogue from the teacher and peer reviewers appears in this window.		Discussion Refresh Clear History	× Send
Comments Library	l ve the test, my dad gave me a lug and a ki pect from other adult drivers, because they	ss, wishing me success on one of the many events that would could actually classify me as know the amount of responsibility that comes along with having a license. But, this individu	a teenager. Fro ual attention has	No records to display.	
be given more fixedom and respect from parents and other adults. The ternage years valued be a time for going out and having a grant time their children. The old clicker "Do surso others and they will do units y that in high school they particle hereby. This is the case of my inter bickes the cost of the house when the was only a fixedom in bids or bickes they are started beauty.	or one thing, parents and other adults were living life to the fullest before we progress io out applies to the fluctuating relationships to Taris, who will turn 30 years old this Decem- hood. Same she was wild during her teenage	once temagers themselves. Having a forme to drive also represents trenagers in ways, be into the stages of additionally are the future addits of the world, and what we have hear there addits and the addetsceness of this world. Everyone the age of 20 and older was a beer. In high school, the was wild, baseally out of control, to Chris and are future would a screen. Just own that doe is an addit- due has been moried for 5 years, has 2 clubbee. In for	ecause there are ned we will pas teenager once, r argue about that one 2 cars, and	ट C ८६ द सर्वे इ. स. सालकार अगर अलग का जार फ सालन (जासकार कि सालकार) अगर अगर अगर का जार का सालन (जासकार	r 1999

Figure 7. Personalized feedback screen

3. The Present Study

3.1 Research Questions

RQ1: Is there a significant statistical difference indicating a positive effect on improving ESL undergraduate students' writing skill following the use of the Criterion® system?

RQ2: What areas of the writing construct appear to improve with the use of the Criterion® system and which areas appear to be unaffected?

3.2 The Context

The undergraduate writing course under study seeks to provide learners with the writing skills necessary to compose a variety of text types in an acceptable to high proficient level. Data was collected through ten different essays submitted by sixty-one EFL undergraduate Saudi female students enrolled in the English Language department between January 2015 and September 2015. The students are provided with ten different essay prompts. The prompts for these ten essays include a variety of topics such as technology, loneliness, wisdom, education, media and team building. A total of 610 essays were submitted and collected automatically via the Criterion® system. Holistic and trait scores as well as automated feedback were assigned for each submitted essay. The students are homogenous in terms of age, gender and educational background but heterogeneous in terms of their English proficiency levels (based on progress reports).Students initially were given a preliminary tutorial session illustrating the system's basic functions.

3.3 Methodological Design

In order to answer the first question posed in this study, a repeated measure design study was administered to identify the effect weight of utilizing automated feedback on students' overall writing quality. A repeated measure design is used in experiments where participants are given more than one treatment and each participant is measured two or more times based on the dependent variable. First, a test of significance, the t-test, was performed in this study to determine if there is a significant variance between the students' pre-test scores (Condition A- the students' first submitted essays) and the students' post-test scores (Condition B- the students' last submitted essay). The matched pair's t-testis used in experiments where two scores, grades or quantities are taken for each participant. It is typically used in studies with before-treatment and after-treatment measurements. Second, a one-way repeated measures ANOVA test was applied to the data in order to assess statistically significant variance in mean scores over the three conditions of treatment. To answer the second research question, students' committed errors were accumulated according to the different writing constructs (grammar, usage, mechanics and style) and compared using descriptive quantitative means.

3.4 Data Analysis

3.4.1 Research Question One

Table 2 reports the descriptive statistics for the students' first performance (condition A), while Table 3 illustrates the same information for their last performance (condition B). As shown below, with a mean of (3.9344) and standard deviation (SD) of (1.223), condition B demonstrates a logical slide in performance compared to condition B with a mean of (2.2787) and SD of (0.9684). The SD was (Hi = 4.00, Low = 1.00) in condition A and (Hi = 6.00 Low = 1.00) in condition B. The median was (2.00) in condition A and (4.00) in condition B, a number that demonstrates a noticeable improvement in students' general performance. The matched-pairs t-test was applied to the data, revealing a statistical difference between the two conditions.

Table 2. Descriptive statistics of condition A

	Des	criptive Statistics					
	Mean	Standard Deviation	Median	Variance	Ν	Sum	Y-Squared
		0.9684					
Group A	2.2787	Hi = 4.00	2.00	0.9377	61	139	373
		Low = 1.00					

Table 3. Descriptive statistics of condition B

	Descri	ptive Statistics					
	Mean	Standard Deviation	Median	Variance	Ν	Sum	Y-Squared
		1.223					
Group B	3.9344	Hi = 6.00	4.00	1.4956	61	240	1034
		Low = 1.00					

Table 4 illustrates the results of the t-test to determine if the variance between the students' initial and last writing performance was significant. The t-statistic was significant at the .05 critical alpha level, t (120) = -8.29, p < .05. Statistical results show that condition A is significantly different from condition B, and we are 95% confident that the mean difference lies between 0.6791 and 2.6324.

Descriptive Statistics			
	Mean	Standard Deviation	n N
Group A	2.2787	0.9684	61
Group B	3.9344	1.223	61
Independent Samples t-T	est		
t-Statistic	-8.29		Result
Degrees of Freedom	120		Reject the null hypothesis.
Critical Value	1.9799		Conclusion
95% Confidence	[0.6791, 2.6324]		Group A is significantly different from Group B, t (120) = -8.29, p
Interval			< .05. We are 95% confident that the mean difference lies between
			0.6791 and 2.6324.
t-Value	-8.2895		
Degrees of Freedom	114.0073		
Two-Tailed p-Value	< 0.0001		
95% Confidence	[-2.0514, -1.26]		
Intervals			

Table 4. Results of the t-test

To trace improvement in student performance over the three measurement times (Condition A, Condition B, and Condition C), a one-way ANOVA test was administered. Table 5 below shows the means and the variance of the three conditions.

Analysis of Variance (One-W	Vay)					
Summary						
Groups	Sample Size	Sum	Mean	Variance		
Condition A	61	140.	2.29508	0.94481		
Condition B	61	202.	3.31148	0.9847		
Condition C	61	239.	3.91803	1.34317		
ANOVA						
Source of Variation	SS	df	MS	F	p-level	F crit.
Between Groups	82.04372	2	41.02186	37.60394	0.%	3.04615
Within Groups	196.36066	180	1.09089			
Total	278.40437	182				

Table 5. Results of the One-Way repeated measures ANOVA test

The ANOVA analysis reveals three important results. First, the analysis includes an F ratio of (37.60394) (p < 0.05), indicating that there is a significant variance between the three conditions in terms of their overall scores. Second, the ANOVA demonstrates significant improvement. Table 5 shows that the sum of the scores was (140) in condition A, (202) in condition B and (239) in condition C. These figures show a gradual improvement in the students' overall performance through the use of the Criterion® system. Third, the variance in condition A was (0.94481), (0.9847) in condition B and (1.34317) in condition C. This variance indicates that the students had begun to show noticeable variations in regards to their general writing improvement in the second phase of the implementation. The source of the variance can be explained by investigating the kinds of errors that were eliminated in students' submitted essays and when this self-correction began, a topic covered in the second research question.

3.4.2 Research Question Two

The second research question posed in this study is concerned with identifying areas within the writing construct that improved through the use of the Criterion® as well as those that did not. Students' errors for all submitted essays have been tallied and summarized in the following figure.



Figure 8. The number of student errors in grammar, usage, mechanics and style

As shown in Figure 1, errors in style ranked the highest, followed by mechanical errors and errors in usage. Grammatical errors were the least common. The following figure illustrates the distribution of these errors across the ten submitted essays.



Figure 9. The distribution of errors among the students' ten submitted essays

As shown in Figure 1, mechanical errors were the most common among students writing their first essay (1186) followed by errors of style (1026). By comparison, there were far fewer mistakes in grammar (212) and usage (266). In students' fifth submitted essay, both mechanical and stylistic errors significantly decreased. Mechanical errors continued to be eliminated until the students' last submission, indicating a significant improvement with only (146) in the students' tenth submitted essay. However, stylistic errors show only moderate improvement between the fifth and tenth essay. Apparently, no significant improvement occurred with regards to grammatical errors in the fifth submitted essay. Moreover, errors of usage increased from (266) to (369) and eventually decreased to only (222).

4. Discussion and Implications

The current paper has aimed to fill the gap in AEE research in the second language context by presenting a longitudinal study investigating the effects of AEE tools on ESL undergraduate students' writing. It has quantified students' levels of improvement and identified areas of the writing construct that appeared to improve and those that did not. The results obtained regarding the first research question are consistent with the inclusive research indicating the critical influence of both automated scoring and automated feedback on writing improvement (Shermis et al., 2008; Wilson et al., 2014; Kellogg et al., 2010). Using a repeated- measure approach measuring improvement across ten consecutive submissions, the research demonstrates that the Criterion® system is beneficial to students' overall writing improvement. The mean of the students' scores on their first submission was (2.28), and increased significantly in the last submission at (3.93), as illustrated in Table 4 above. The mean difference between the students' first overall performance and their last one was statistically significant (between 0.68 and 2.63). The total number of errors diagnosed and detected by the Criterion® system in the students' first submission was (2.2600), which declined significantly to only (1.226).

This significant improvement results from the variety of Criterion® tools that assist both the process of automated scoring and individualized automated feedback. The Criterion® provides instructor with an archive of essay topics with a variety of levels and modes. The instructor can also create his own topics. It also offers online tracking of learners' portfolios including submitted essay, progress and an overall all evaluation of writing skills. Furthermore, the instructor can even customize the system instructions to best select level-appropriate writing resources and feedback. The Criterion® also projects summary class reports that analyze the learner's overall progress and patterns of errors. The availability of all these options results in more writing assignments assigned to learners involving more opportunities to practice writing; more time to assist learners in acquiring the higher-order skills of writing; and more effective interaction between instructor and learners. Moreover, the Criterion® system provides learners with the core features that are the most essential in improving non-native English learners' writing skills: prewriting strategies in the drafting process, friendly, easy-to-use online planning tools, immediate feedback, opportunities to make revisions and content-related instructor feedback (e.g., Covill, 1997; Etchison, 1989; Fitzgerald, 1987). The mechanism underlying the Criterion® system is in accordance with

Anderson's (1985) approach of language learning. Anderson's approach is stated in three consecutive stages: construction, transformation and execution. The stage of construction involves planning for the task by brainstorming, using mind-mapping strategies and outline. The transformation stage is when language rules are practiced to transform intended meanings into the form of writing, composing and revising. The execution stage is related to the physical process of composing the text.

However, the students' writing improved mostly in mechanics after using the Criterion® system. These types of errors significantly declined. Errors related to style constitute the majority of students' committed errors throughout the study, and show moderate improvement compared to improvement in mechanics. Similarly, areas of grammar and usage showed only moderate improvement. These results align with Lee et al. (2009) who reported that there were no significant effects of applying a web-based critiquing tool with adult EFL learners in regard to both the content and the organization of students' writing. The results of the present study align also with the argument of Warschauer & Grimes (2008) who did not support the use of the Criterion® system nor My Access tools in first language context, claiming that the effectiveness of such automated tools reside only at the level of mechanics, including punctuation, spelling and grammar. The results of Research Q2 in the present study confirms what Warschauer & Grimes (2008, p. 4) indicated in their study that AEE systems "remain relatively error-prone and insensitive to individual learners' skills and needs". Like any other technological tool in educational context, AEE systems should be implemented with an awareness of its benefits and flaws.

The findings obtained from the present study suggest a number of issues that should be taken into consideration when implementing AEE systems in educational contexts. First, a distinction has to be drawn between such systems' scoring and supporting functions. The Criterion® system, like many other AEE systems, has proven efficient for the purposes of assessing, certifying and classifying students in terms of their writing proficiency levels. However, caution is advised when utilizing the Criterion® system for instructional purposes and assisting students during drafting, composing and revising. The Criterion® system cannot replace human review that encourages students to be involved cognitively in the writing process. Second, the Criterion® system has great potential for tracking progress and generating individualized student portfolios, including areas of strength and weakness. Portfolios have been depicted as crucial pedagogical tools integrating both assessment and instruction in the context of learning and teaching writing. Hamp-Lyons (1994) generated numeral portfolios obtained through the Criterion® system to facilitate instruction and assessment. Third, different language proficiency levels in the L2 should be taken into account while utilizing the Criterion® system in instructional feedback. Beginner students will not benefit from the feedback given by the system nor suggestions for editing and correcting. Future research is needed to target the correlation between proficiency levels and areas of improvement.

In the present study, the utilization of the Criterion® system is allowing the complex analysis of writing tasks for both assessment and instruction. It provides an in-depth understanding of how AEE should be conceptualized and operationalized in the context of writing assessment and instruction. However, AEE research technology continues to evolve. This changing environment calls on all AEE stakeholders to become involved in shaping the future development of AEE technology as current features could still be improved. Perhaps additional noteworthy studies are needed to address the integration of AEE technology into writing curricula. Instructions and guidelines on how to integrate AEE systems into writing contexts also are needed for those seeking to design materials for writing classes and for instructors seeking to apply AEE systems in their classrooms.

References

- Ananiadou, K., & Claro, M. (2009). 21st century skills and competences for new millennium learners in OECD countries. OECD Education Working Pages, 41. http://dx.doi.org/10.1787/218525261154
- Anderson, J. (1985). Cognitive psychology and its implications. New York: W. H. Freeman.
- Attali, Y. (2004). *Exploring the feedback and revision features of Criterion*. Paper presented at the National Council on Measurement in Education (NCME), San Diego, CA.
- Bejar, I. I. (2011). A validity-based approach to quality control and assurance of automated scoring. Assessment in Education: Principles, Policy & Practice, 18(3), 319-341. http://dx.doi.org/10.1080/0969594X.2011.555329.
- Bejar, I. I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice*, 31(3), 2-9. http://dx.doi.org/10.1111/j.1745-3992.2012.00238.x.
- Black, P., & William, D. (1998). Inside the black box: Raising standards through classroom assessment [Electronic version]. *Phi Delta Kappan, 80,* 139-148. Retrieved from

http://www.pdkintl.org/kappan/kbla9810.htm

- Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The Criterion online writing service. *American Association for Artificial Intelligence*, 25(3), 27-36.
- Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., & Harris M. D. (1998). Automated scoring using a hybrid feature identification technique. In *proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics* (pp. 206-210). East Stroudsburg, PA: Association for Computational Linguistics.
- Burstein, J., Marcu, D., & Knight, K. (2003). Finding the WRITE stuff: Automatic identification of discourse structure in student essays. *IEEE Intelligent Systems: Special Issue on Natural Language Processing*, 18(1), 32-39. http://dx.doi.org/10.1109/MIS.2003.1179191
- Cohen, Y., Ben-Simon, A., & Hovav, M. (2003). *The effect of specific language features on the complexity of systems for automated essay scoring*. Paper presented at the 29th Annual Conference of the International Association for Educational Assessment, Manchester, UK.
- Common Core State Standards Initiative. (2012). Preparing America's students for college & career. Retrieved from www.corestandards.org
- Covill, A. (1997). Students' revision practices and attitudes in response to surface-related feedback as compared to content-related feedback on their writing. *Dissertation Abstracts International*, 58. (UMI No. 9716828).
- Ebyary, K. E., & Windeatt, S. (2010). The impact of computer-based feedback on students' written work. *International Journal of English Studies*, 10(2), 121-142.
- Etchison, C. (1989). Word processing: A helpful tool for basic writers. *Computers and Composition*, 6(2), 33-43. http://dx.doi.org/10.1016/S8755-4615(89)80013-1
- Ferris, D. (2011). Treatment of errors in second language student writing. Ann Arbor, MI: University of Michigan Press. http://dx.doi.org/10.3998/mpub.2173290
- Fitzgerald, J. (1987). Research on revision in writing. *Review of Educational Research*, 57(4), 481-506. http://dx.doi.org/10.3102/00346543057004481
- Franzke, M., Kintsch, E., Caccamise, D., Johnson, N., & Dooley, S. (2005). Summary street: Computer support for comprehension and writing. *Journal of Educational Computing Research*, 33, 53-80. http://dx.doi.org/10.2190/DH8F-QJWM-J457-FQVB
- Graham, S., Harris, K. R., & Hebert, M. (2011).*Informing writing: The benefits of formative assessment*. Washington, DC: Alliance for Excellence in Education.
- Grimes, D., & Warschauer, M. (2008). Learning with laptops: A multi-method case study. *Journal of Educational Computing Research*, 38(3), 305-332. http://dx.doi.org/10.2190/EC.38.3.d.
- Hamp-Lyons, L., & Kroll, B. (1996). Issues in ESL writing assessment: An overview. CollegeESL, 6(1), 52-72.
- Hattie, J., & Timperely, H. (2007). The power of feedback. *Review of Educational Research*, 77, 81-112. http://dx.doi.org/10.3102/003465430298487
- Kellogg, R. T., Whiteford, A. P., & Quinlan, T. (2010). Does automated feedback help students to write? *Journal of Educational Computing Research*, 42, 173-196. http://dx.doi.org/10.2190/EC.42.2.c
- Kulik, J. A. (2003). Effects of using instructional technology in elementary and secondary schools: What controlled evaluation studies say. Arlington, VA: SRI.
- Lee, C., Wong, K., Cheung, W., & Lee, F. (2009). Web-based essay critiquing system and EFL students' writing: A quantitative and qualitative investigation. *Computer Assisted Language Learning*, 22, 57-72. http://dx.doi.org/10.1080/09588220802613807
- Omaggio, H. A. (1993). Teaching language in context. Boston: Heinle & Heinle.
- Shermis, M. D., Wilson G. C., & Diao, Y. (2008). *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, NY.
- Shermis, M., & Burstein, J. (2013). *Handbook of automated essay evaluation*. New York and London: Routledge.
- Silva, T. (1993). Towards and understanding of distinct nature of L2 writing. The ESL research and its

implications. TESOL Quarterly, 27(4), 657-677. http://dx.doi.org/10.2307/3587400

- Valenti, S., Neri, F., &Cucchiarelli, A. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education*, 2, 319-330.
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.
- Wang, J., & Brown, M. S. (2007). Automated essay scoring versus human scoring: A comparative study. *The Journal of Technology, Learning, and Assessment*, 6(2), 1-28.
- Wang, Y. J. (2011). Exploring the effect of using automated writing evaluation in Taiwanese EFL students' writing. Unpublished Master's thesis. I-Shou University, Taiwan.
- Warschauer, M., & Grimes, G. (2008). Automated writing assessment in the classrooms. *Pedagogies: An International Journal*, *3*, 22-36. http://dx.doi.org/10.1080/15544800701771580.
- Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. Language Teaching Research, 10(2), 157-180. http://dx.doi.org/10.1191/1362168806lr190oa
- Wilson, J., Olinghouse N. G., & Andrada, G. N. (2014). Does automated feedback improve writing quality? *Learning Disabilities: A Contemporary Journal, 12*, 93-118.
- Zamel, V. (1985). Responding to student writing. *TESOL Quarterly*, 19, 79-102. http://dx.doi.org/10.2307/3586773

Notes

Note 1. A holistic score gives a general indicator of the text's writing based on common sets of evaluation criteria. A trait score, on the other hand, gives a single indicator of one of the core features of writing (i.e., organization, grammar, mechanics, etc.). It therefore allows the provision of suggestions for areas of weakness and strength.

Note 2. Condition A represents the students' first submitted essay. Condition B represents the students' fourth essay. Condition C represents the students' last essay.

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/).