REPUBLIC OF TURKEY
YILDIZ TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF SOCIAL SCIENCES
DEPARTMENT OF FOREIGN LANGUAGES EDUCATION
ENGLISH LANGUAGE TEACHING PROGRAM

MASTER'S THESIS

THE IMPACT OF AUTOMATED WRITING EVALUATION AND COMBINED AUTOMATED – TEACHER FEEDBACK ON EFL LEARNERS' WRITING PERFORMANCE

MERVE BAKO
21708003

THESIS ADVISOR
ASSOC. PROF. DR. SUZAN KAVANOZ

2023

REPUBLIC OF TURKEY
YILDIZ TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF SOCIAL SCIENCES
DEPARTMENT OF FOREIGN LANGUAGES EDUCATION
ENGLISH LANGUAGE TEACHING PROGRAM

MASTER'S THESIS

THE IMPACT OF AUTOMATED WRITING EVALUATION AND COMBINED AUTOMATED – TEACHER FEEDBACK ON EFL LEARNERS' WRITING PERFORMANCE

MERVE BAKO
21708003
ORCID NO: 0000-0002-2283-1150

THESIS ADVISOR
ASSOC. PROF. DR. SUZAN KAVANOZ
ORCID NO: 0000-0001-7458-4684

JUNE 2023

Merve Bako tarafından hazırlanan “The Impact of Automated Writing Evaluation and Combined Automated – Teacher Feedback on EFL Learners’ Writing Performance” başlıklı çalışma, **12/06/2023** tarihinde yapılan savunma sınavı sonucunda oybirliği ile başarılı bulunmuş ve jürimiz tarafından İngiliz Dili Eğitimi Ana Bilim Dalı Yabancı Diller Eğitimi Programında **YÜKSEK LİSANS** tezi olarak kabul edilmiştir.

Danışman	**İmza**
Doç. Dr. Suzan Kavanoz	……

Jüri Üyeleri	**İmza**
Dr. Öğr. Üyesi, Burcu Ünal	……
Dr. Öğr. Üyesi, Mustafa Polat	……

ABSTRACT

THE IMPACT OF AUTOMATED WRITING EVALUATION AND COMBINED AUTOMATED – TEACHER FEEDBACK ON EFL LEARNERS' WRITING PERFORMANCE

Over the past few years, writing proficiency has become an essential component of 21st-century skills. Writing skills in first, second, and foreign languages have continuously been assessed and graded by teachers or professional raters trained solely to be objective and consistent. Some inconsistencies in teacher assessment can still occur because of the human element. To address this issue, some tools have been developed to assess and evaluate student writing and incorporate automated feedback, which can be used in writing classrooms. This study uses a naturalistic classroom-based method to investigate how the use of an AWE tool called Criterion as part of writing instruction and receiving combined automated teacher feedback affect the students' writing performance and how they view the program's efficacy in improving their writing. In the current study, a mixed methods research design was utilized, involving 38 EFL students from preparatory classes in two intact groups. Quantitative and qualitative data collection instruments were employed to gather the necessary data. The results illustrated that using an AWE tool (Criterion) and receiving combined automated–teacher feedback did not have a significant impact on the writing performance of the students in the experimental group. However, the EFL learners in the experimental group improved significantly in their writing performance before and after the intervention. One-on-one interviews revealed a positive attitude towards the use of Criterion and receiving combined automated–teacher feedback. The students felt motivated and engaged in more writing, thanks to the immediate feedback and the chance to write multiple drafts provided by Criterion. It was expressed that the whole process helped the students to learn how to write a persuasive essay properly.

Keywords: automated writing evaluation (AWE), automated essay scoring (AES), automated feedback, teacher feedback, combined automated–teacher feedback, writing assessment, e-rater, Criterion

ÖZET

OTOMATİK YAZMA DEĞERLENDİRME VE OTOMATİK - ÖĞRETMEN GERİ BİLDİRİMİNİN YABANCI DİL OLARAK İNGİLİZCE ÖĞRENİCİLERİNİN YAZMA PERFORMANSI ÜZERİNDEKİ ETKİSİ

Son birkaç yılda, kompozisyon yazma konusunda yetkin olmak, 21. yüzyıl becerilerinin temel bir bileşeni haline gelmiştir. Ana dil ve yabancı dillerdeki yazma becerileri, şimdiye değin yalnızca objektif ve tutarlı olmak üzere eğitilmiş öğretmenler veya profesyonel puanlayıcılar tarafından sürekli olarak değerlendirilmiş ve puanlanmıştır. Öğretmen değerlendirmesinde, insan faktörü nedeniyle bazı tutarsızlıklar ortaya çıkmaktadır. Bu sorunu çözebilmek adına, öğrenci kompozisyonlarını ölçmek ve değerlendirmek için bazı programlar geliştirilmiştir. Bu otomatik kompozisyon değerlendirme araçları, otomatik geri bildirim içerir ve bu araçlar kompozisyon yazma derslerinde kullanılabilir. Bu çalışma, kompozisyon yazma dersinin bir parçası olarak Criterion adlı bir otomatik kompozisyon değerlendirme aracının kullanımının ve öğretmen geri bildirimine ek olarak otomatik geri bildirim de almanın öğrencilerin kompozisyon yazma performansını nasıl etkilediğini ve programın, öğrenciler tarafından yazılan kompozisyonlar üzerinde nasıl bir etkisi olduğunu görmek için doğal sınıf tabanlı bir yöntem kullanır. Mevcut çalışmada, hazırlık sınıflarında öğrenim gören 38 İngilizce'yi yabancı dil olarak öğrenen öğrencinin iki sınıfta toplandığı karma yöntem araştırma deseni kullanılmıştır. Gerekli verileri toplamak için hem nicel hem de nitel veri toplama araçlarından faydalanılmıştır. Sonuçlar, bir otomatik kompozisyon değerlendirme aracı kullanmanın ve otomatik geri bildirim ve öğretmen geri bildirimi almanın, deney grubu öğrencilerinin yazma performansı üzerinde anlamlı bir etkiye sahip olmadığını göstermişitr. Bununla birlikte, deney grubundaki İngilizce'yi yabancı dil olarak öğrenen öğrencilerin yazma performansları uygulama öncesine göre önemli ölçüde artmıştır. Öğrencilerle yapılan bire bir görüşmeler, Criterion kullanımına ve otomatik geri bildirim ve öğretmen geri bildirimi almaya yönelik olumlu bir tutum ortaya koymuştur. Criterion'ın sağladığı hızlı geri bildirim ve birden fazla taslak yazma şansı sayesinde öğrencilerin motive oldukları ve daha fazla metin üretebildiklerini bulgusuna ulaşılmıştır. Tüm sürecin, öğrencilerin kompozisyon yazmada yetkin hale gelmelerine yardımcı olduğu da ifade edilmiştir.

Anahtar Kelimeler: Otomatik kompozisyon değerlendirme, otomatik puanlayıcı, otomatik geri bildirim, öğretmen geri bildirimi, kompozisyon değerlendirme, e-rater, Criterion

ACKNOWLEDGEMENTS

I would like to extend my deepest gratitude to my advisor, Assoc. Prof. Dr. Suzan Kavanoz for her unwavering support, invaluable guidance, and continuous encouragement throughout this research journey. Her expertise, insights, and constructive feedback have been instrumental in shaping the direction and quality of this thesis.

I am immensely grateful to the faculty members of Yıldız Technical University, Foreign Languages Education Department, whose knowledge, teachings, and mentorship have greatly influenced my academic and research endeavors. Their commitment to excellence and passion for their respective fields have inspired me to push the boundaries of my knowledge and skills.

I would also like to thank my colleagues at Yıldız Technical University, School of Foreign Languages, for their support and collaboration during my research. In particular, I would like to thank Sibel Atmaca, Burçin Engürel, Zeynep Kandemir, Yağız Dikkatli, Kürşat Gültekin, and Göksel Kaya for their help with evaluating the essays.

I am grateful to my friends and classmates for their unwavering support, encouragement, and countless discussions that have broadened my perspectives and enhanced the quality of my work. Their camaraderie and shared experiences have made this academic journey more enjoyable and memorable.

I would like to extend my sincere gratitude to the participants of this study, without whom this research would not have been possible. Their willingness to share their experiences, time, and insights has provided invaluable insights and enriched the findings of this thesis. Thank you for your time and contribution.

I am grateful to Educational Testing Service (ETS) for providing me with complimentary licenses to use Criterion Online Writing Evaluation Service for research purposes.

Lastly, I would like to express my heartfelt appreciation to my husband, Reşat Bako, for his unconditional love, understanding, and encouragement throughout my academic pursuits. His unwavering support and belief in my abilities have been the driving force behind my accomplishments.

In conclusion, I would like to acknowledge the collective effort and support of all those who have contributed to the successful completion of this master's thesis. Their contributions, whether large or small, have left an indelible mark on this research, and for that, I am truly grateful.

Merve BAKO

June, 2023; İstanbul

TABLE OF CONTENTS

ABSTRACT iii
ÖZET iv
ACKNOWLEDGEMENTS v
TABLE OF CONTENTS vi
LIST OF TABLES ix
LIST OF FIGURES xi
LIST OF ABBREVIATIONS xii
1. INTRODUCTION 1
1.1. Statement of the Problem 2
1.2. Purpose of the Study 3
1.3. Significance of the Study 4
1.4. Research Questions 5
1.5. Definitions of the Key Terms 5
2. LITERATURE REVIEW 7
2.1. Writing Construct 7
2.2. Assessment of Writing 7
2.2.1. Holistic and Trait Scoring Approaches 8
2.3. Human Scoring and Feedback 9
2.3.1. Benefits of Human Scoring and Feedback 10
2.3.2. Drawbacks of Human Scoring and Feedback 10
2.4. History of Automated Writing Scoring & Automated Writing Evaluation 13
2.4.1. Traditional AES/AWE Models 15
2.4.1.1. Trins and Proxes by PEG 16
2.4.1.2. Latent Semantic Analysis (LSA) 16
2.4.1.3. Natural Language Processing (NLP) 17
2.5. Different AES/AWE Systems 18

2.5.1. Project Essay Grade (PEG) 19
2.5.2. Intelligent Essay Assessor (IEA) 20
2.5.3. IntelliMetric and MY Access! 20
2.5.4. E-rater and Criterion 21
2.6. Automated Scoring and Feedback 23
2.6.1. Validity and Reliability of Automated Essay Scoring 24
2.6.2. Automated Feedback 25
2.6.3. Human Scoring and Feedback vs. Automated Scoring and Feedback 26
2.7. The Impact of AWE in Classroom 28
2.7.1. Positive Effects of AWE Systems 28
2.7.1.1. Improvement in Writing Skills 28
2.7.1.2. Students' Attitude 30
2.7.1.3. Instructors' Attitude 31
2.7.2. Criticism towards AWE Systems 32
2.8. Combined Teacher and AWE Feedback 34
3. METHODOLOGY 36
3.1. Research Design 36
3.2. The Setting of the Study 38
3.3. Participants 39
3.3.1. Students 39
3.3.2. The Instructor and the Raters 42
3.4. Data Collection Instruments 43
3.4.1. Placement test 43
3.4.2. Pre-test 43
3.4.3. Post-test 44
3.4.4. Weekly Writing Practices 46
3.4.5. Essay Evaluation Rubric 46

3.4.6. One-on-one Interviews 46
3.5. Data Collection Procedure 47
3.5.1. Intervention Procedure Each Week 49
3.5.1.1. Intervention and Feedback Procedure for the Control Group 49
3.5.1.2. Intervention and Feedback Procedure for the Experimental Group .. 51
3.6. Data Analysis 59
3.6.1. Quantitative Data Analysis 59
3.6.2. Qualitative Data Analysis 60
4. RESULTS 62
4.1. Quantitative Data Analysis Results 62
4.1.1. The Results of the Pre-test and Post-test 63
4.2. Qualitative Data Analysis Results 68
4.2.1. The Results of the One-on-one Interviews 68
5. CONCLUSION 83
5.1. Discussion of the Findings 83
5.1.1. Discussion on the Findings for RQ1 83
5.1.1.1. Difference between groups 83
5.1.1.2. Difference within groups 86
5.1.2. Discussion on the Findings for RQ2 87
5.2. Conclusion 97
5.3. Pedagogical Implications 99
5.4. Limitations of the Study 100
5.5. Recommendations for Further Research 101
REFERENCES / BIBLIOGRAPHY 103
APPENDICES 115

LIST OF TABLES

Table 1. Different Dimensions of Essay Quality 14

Table 2. Different AES/AWE Systems 18

Table 3. The Departments of the Students 40

Table 4. Descriptive Statistics of the Placement Test 41

Table 5. Mann-Whitney U Test to Compare the Placement Test Results of the Experimental and the Control Groups 41

Table 6. List of Categories and Subcategories of Criterion Feedback 53

Table 7. Study Plan 54

Table 8. Descriptive Statistics for Pre-tests of Both Groups 63

Table 9. Mann-Whitney U Test to Compare the Pre-test Scores of the Experimental and the Control Groups 63

Table 10. Wilcoxon Signed Rank Test to Compare the Pre-test and the Post-test Scores of the Experimental Group 64

Table 11. Wilcoxon Signed Rank Test to Compare the Pre-test and the Post-test Scores of the Control Group 65

Table 12. Descriptive Statistics for Post-tests of Both Groups 65

Table 13. Mann-Whitney U Test to Compare the Post-test Scores of the Experimental and the Control Groups 66

Table 14. Descriptive Statistics for Gain Scores of Both Groups 67

Table 15. Mann-Whitney U Test to Compare the Gain Scores of the Experimental and the Control Groups 67

Table 16. Reflections about the Accuracy of the Scores Given by Criterion 69

Table 17. Students' Strategies to Increase their Scores on Criterion 70

Table 18. Students' Preferred Feedback Type 72

Table 19. Reflections on Different Types of Feedback 72

Table 20. Reflections on Improved Writing Skills 74

Table 21. The Strengths and Weaknesses of Criterion 76

Table 22. Motivational Factors Related to Criterion 78

Table 23. Suggestions Related to the Whole Process or the Use of Criterion 81

LIST OF FIGURES

Figure 1. Research Design 37

Figure 2. Criterion Holistic and Trait Level Scores 58

Figure 3. Criterion Feedback on Different Categories 58

Figure 4. Highlighted Errors and Advisory Messages 58

Figure 5. Teacher Feedback on Criterion 59

LIST OF ABBREVIATIONS

AEE	: Automated Essay Evaluation
AES	: Automated Essay Scoring
AI	: Artificial Intelligence
AWE	: Automated Writing Evaluation
CCSI	: Common Core State Standards Initiative
EFL	: English as a Foreign Language
ESL	: English as a Second Language
ETS	: Educational Testing Service
EPE	: English Proficiency Exam
GMAT	: Graduate Management Admission Test
GRE	: Graduate Record Examination
IEA	: Intelligent Essay Assessor
IELTS	: International English Language Testing System
L2	: Second Language
LSA	: Latent Semantic Analysis
MI	: Measurement Incorporated
NLP	: Natural Language Processing
PEG	: Project Essay Grade
PKT	: Pearson Knowledge Technologies
PTE	: Pearson Test of English
TOEFL	: Test of English as a Foreign Language

1. INTRODUCTION

Conducting essay writing tests has a lengthy history. Even around 2,100 years ago, the Mandarin government of China held written civil service examinations (Page & Petersen, 1995) to be able to make sure that people who worked for the government had this skill.

Undeniably, writing is a significant academic skill that takes much time and energy both for teachers and students to improve. Many teachers believe that writing is a productive skill that is hardest to acquire and, consequently, to teach because of the attention required to accuracy due to its intricate communication process (Hyland & Hyland, 2006).

For centuries, students have been making a sincere effort to enhance their writing abilities in both their first language and second language. Likewise, teachers are also dedicatedly investing their time and effort to aid their students' writing progress by assigning homework, grading their work, and giving constructive feedback. This rather significant skill has been taught and evaluated by the teachers.

Although grading papers and giving feedback is part of the teaching occupation, the teacher’s time allocated for each student is limited, especially when there are a large number of students. Moreover, teacher feedback may not always be consistent due to the inherent fallibility of human beings. These two concerns related to teachers and the importance of the improvement of the writing skill combined led researchers to find a solution that addresses this problem: how the students get enough and consistent feedback to improve their writing skills.

Automated Essay Scoring (AES) and Automated Writing Evaluation (AWE) programs have been developed as an alternative to human scoring and feedback. There are two components of AWE systems. While the first one is the scoring engine that plays a crucial role and is responsible for producing automated scores using methods like artificial intelligence, natural language processing, and latent semantic analysis, the

second one is the written feedback that can be used as part of the classroom instruction of the writing skill (Stevenson & Phakiti, 2014).

However, there is a certain level of disagreement among researchers and educators about the effectiveness of the AWE software. While supporters believe that AWE is a valuable tool for enhancing students' writing skills, how well these automated tools can assess creative writing continues to be questioned by critics.

Automated Essay Scoring (AES) has been increasingly adopted by educators as a summative assessment tool, and AES also offers assistive writing resources to teachers and students as formative assessment. Over the past few decades, developing different Automated Writing Evaluation (AWE) systems that utilize advanced language processing technologies has widened the array of feedback resources accessible to learners for their language learning. AWE programs offer a multitude of advantages over traditional human evaluation, including the speed of online feedback (Dikli, 2006), the provision of diagnostic feedback and overall grades on the writings of the learners (Shermis & Burstein, 2003), opportunities for multiple revisions (Warschauer & Ware, 2006), and favorable impacts on learner autonomy (Chen & Cheng, 2008).

1.1. Statement of the Problem

Human raters are inconsistent in the scores and the feedback they give (Brimi, 2011), and this can be attributed to a variety of factors such as interpreting the rubrics differently (Attali, 2013), subjectivity (Shin & Gierl, 2021), reader bias (Saal, Downey & Lahey, 1980), which can occur as a result of reader's characteristics, psychology, and rating conditions (Attali, 2016; Barkaoui, 2010a; Barkaoui, 2010b; Braun, 1988; Engelhard, 1994; Hoskins & Wilson, 2001; Quellmalz, 1980; Rao & Li, 2017). Moreover, teachers find it challenging to spot every mistake when it happens, leading to inconsistency in the feedback conveyed by the teacher, causing confusion for the students (Ranalli, Link, & Chukharev-Hudilainen, 2017).

In addition to inconsistency, human scoring is not time-effective when instructors' busy schedules and the number of students in their classes are considered (Huang & Renandya, 2020). Trying to provide personalized and prompt feedback and answer students' questions related to the evaluation process results in a significant workload for teachers who are required to rectify many written assignments submitted by their

students (Chapelle, Cotos, & Lee, 2015; Grimes & Warschauer, 2010; Warschauer & Grimes, 2008; Wilson & Czik, 2016). As a result, the feedback becomes a summative score, delayed, and fewer assignments are given (Dikli, 2006; Huang & Renandya, 2020).

It is suggested that reliability and consistency can be increased through constant training and using multiple raters (Douglas, 2009). However, it is very costly. Adopting Automated Essay Scoring (AES) can tackle these issues by offering timely and accurate scores (Lottridge, Schulz, & Mitzel, 2013; Shin & Gierl, 2021).

Literature shows that many studies have been conducted since Page's (1966) first mention of automated essay scoring. Some studies explored the effectiveness of AWE tools (Attali, 2004; Choi, 2010; Ebyary & Windeat, 2010; Wang, Shang, & Briody, 2013; Wilson & Czik, 2016) with different results because they employed different research designs. Some studies only had experimental groups (Jinla, 2014; Li, Link, & Hegelheimer, 2015; Parra & Calero, 2019; Scharber, Dexter, & Riedel, 2008; Wali & Huijser, 2018), while some studies compared the experimental group that used an AWE tool to a control group in no feedback condition (Cheng, 2017) or a control group that received teacher feedback (Wang et al., 2013). However, the number of studies that examined the impact of using both automated and teacher feedback compared to using only teacher feedback is limited (Choi, 2010; Wilson & Czik, 2016). It is essential to know whether these tools contribute to students' writing abilities.

Moreover, AWE has been frequently scrutinized for its reliability in standardized tests conducted on a large scale (Attali, 2004; Attali & Burstein, 2006; Powers, Burstein, Chodorow, Fowles, & Kukich, 2001; Stevenson, 2016); studies on using AWE as formative assessment in writing classrooms are not many (Chen & Cheng, 2008; Ebyary & Windeat, 2010; Li et al., 2015). That is why more studies are needed in pursuit of proving the efficiency of AWE tools in enhancing learners' writing skills.

1.2. Purpose of the Study

Automated writing evaluation tools seem quite promising to be actively used as part of EFL /ESL writing instruction. They may help the writing instructors in many areas and relieve their paper-grading burden. However, there is still a need to understand whether they can live up to their claimed effectiveness.

The purpose of this study is to explore whether using an AWE tool has an impact on students' writing performance and how a combined approach of automated and teacher feedback affects the writing development of Turkish EFL students at the tertiary level, in comparison to the traditional teacher feedback only model. Additionally, this study aims to investigate the perception of the learners who received combined automated-teacher feedback on their essays and their attitudes towards writing in English.

1.3. Significance of the Study

Every stakeholder in the assessment and teaching of writing to learners of English understands and wants to know more about the benefits and drawbacks of automated writing scoring and evaluation tools. Moreover, there are numerous gaps and recommendations for the future in the literature that made this research inevitable.

It is believed that AWE can help instructors by enabling them to focus on writing abilities at a higher level, such as content and organization, while AWE tools address writing skills at a lower level, such as grammar and mechanics. Therefore, student writing will be positively affected. However, as Link, Mehrzad, and Rahimi (2020) claim, these assertions are not well-supported by evidence, casting doubt on the effect of AWE on instruction and learning.

There are many studies conducted on testing the validity and reliability of AWE systems (e.g., Attali, 2004; Attali & Burstein, 2006; Powers et al., 2001; Shermis & Burstein, 2003, 2013; Stevenson, 2016); however, studies on the efficacy of AWE as an educational tool in writing classrooms are quite few (Chen & Cheng, 2008). Stevenson & Phakiti (2014) discovered that literature paid relatively little attention to how AWE was utilized in the classroom or how it could be successfully incorporated into writing lessons. This study has been conducted to answer this problem.

AWE feedback has the potential to improve writing skills; however, there is currently insufficient research to prove its efficacy for use in L2 writing courses, and further empirical research on how the use of AWE tools in writing classes affects teaching or learning would be priceless (Link et al., 2020).

Studies done on using AWE in writing classes focused only on experimental groups, without including control groups to observe if AWE helped the students or not (e.g., Chen & Cheng, 2008; Jinla, 2014; Li et al., 2015; Parra & Calero, 2019; Wali &

Huijser, 2018). For this reason, this study has a quasi-experimental research design in which there is an experimental group to try the AWE tool and a control group to observe the difference.

Many academics assert that these commercial AWE systems do not adequately serve English language students with diverse linguistic and cultural backgrounds (e.g., Wali & Huijser, 2018), and additional research is needed to assess how well these systems perform in diverse contexts (Almusharraf & Alotaibi, 2022). Therefore, conducting this research with Turkish EFL learners will also investigate this issue.

The participants of the studies conducted on the use of AWE tools were mostly English majors (e.g., Chen & Cheng, 2008; Link et al., 2020; Wang, 2013; Wang et al., 2013). However, in the present study, the learners of English from different majors, mainly engineering, participated and utilized an AWE tool.

Because of the reasons mentioned above, this study bears great significance in terms of the contribution it made to comprehend the efficiency of using AWE tools in writing classes on learners' writing performance.

1.4. Research Questions

1. Do the use of Criterion and receiving combined automated-teacher feedback affect Turkish EFL students' writing performance significantly compared to teacher feedback only?

2. What are the experimental group students' perceptions of combined automated teacher feedback and their attitudes towards using Criterion as part of their writing classes?

1.5. Definitions of the Key Terms

Automated Essay Scoring: "Automated essay scoring (AES) is the ability of computer technology to evaluate and score written prose" (Shermis & Burstein, 2003, p: xiii).

Automated Essay Evaluation: Automated essay evaluation (AEE) can be defined as using computer programs to assess and rate written content. However, it is essential to note that the word "evaluation" in this term acknowledges the potential for the

technology to do more than just assign a score or grade to an essay; it also offers diagnostic feedback (Shermis, Burstein & Bursky, 2013).

Formative Assessment: The term "formative assessment" can be explained as actions carried out by teachers and their students, which involves self-assessment, to obtain data that can be used as feedback to adjust teaching and learning activities (Black & Wiliam, 2010).

Summative Assessment: Summative assessment is “cumulative assessments...that intend to capture what a student has learned, or the quality of the learning, and judge performance against some standards” (National Research Council, p. 25).

2. LITERATURE REVIEW

2.1. Writing Construct

Constructing a text requires both microstructural and macrostructural features. While microstructural features include syntactic complexity, using discourse markers, vocabulary choices, spelling, punctuation, and capitalization, macrostructural features refer to text's general organization, the variety of functional elements it contains, overall coherence, the depth of its subject matter, and/or the efficiency of the communication. While macrostructural writing characteristics are challenging to evaluate, microstructural components of text composition are simpler to operationalize and measure (Tolchinsky, 2016).

The writing construct designed for Automated Evaluation (AES) systems involves features that create a mutual agreement among the things that are desirable in writing, measurable through AES models, and feasible with regard to the level of computational power and intricacy of programming required (Deane, 2013).

2.2. Assessment of Writing

Writing assessment has been created to precisely but plainly measure the linguistic assets that each student brings to their own writing. It has become a crucial instrument in evaluations when English proficiency needs to be determined, in addition to enabling measures of other subjects and skills (Rezaei & Lovorn, 2010).

To be able to measure writing ability, the assessment rubrics are designed based on the characteristics that are immediately noticeable (observed variable) and employ these traits as a placeholder for the writing features that are not immediately obvious (latent variable) (Koskey & Shermis, 2013). The outcome of an IQ test is an example of an observed variable, but the underlying construct being measured, like intelligence, is an example of a latent variable (Shermis et al., 2013).

It is possible to identify three qualities of an essay that are essential to evaluate: the accuracy and thoroughness of the conceptual knowledge it contains, the validity of the

arguments it uses to discuss various topics, and the writing's fluency, elegance, and readability when we think the concept of writing at an abstract level (Foltz, Laham, & Landauer, 1999)

Moreover, teachers use 14 criteria to find these linguistic features in students' writing. Assessment at the genre level concentrates on the social goal, stages, and phases of the work. It concentrates on the field, tenor, and mode of the text at the register level. Lexical, conjunctional, and reference resources are recognized at the discourse level. Spelling, punctuation, and visual presentation are included in the level of graphic characteristics, whereas the diversity and correctness of grammatical resources are assessed at the grammar level. The assessment is therefore conducted top-down, starting with genre and moving on to register, characteristics related to discourse, grammatical structures that create discourse, and graphological features that reflect patterns in writing (Rose, 2016).

2.2.1. Holistic and Trait Scoring Approaches

There are two types of rubrics which are holistic rubrics and trait rubrics. Rubrics based on traits assess the features of writing which are significant to a specific audience and writing objective (Stiggins & Bridgeford, 1983). On the other hand, when a holistic rubric is employed, the evaluator assesses an essay based on its overall quality. In this approach, the evaluator considers multiple criteria simultaneously while assigning a single score. This score reflects the examinee's performance on a combination of characteristics, including the use of words, writing conventions, voice, syntactic variety, grammaticality, and organizational quality (Koskey & Shermis, 2013). When it comes to discourse coherence, holistic scoring guides offer a vague understanding of the language elements that could impact the evaluation of the coherence quality. However, they do not give clear instructions or examples to the evaluator on the particular language components that lead to better or worse coherence quality (Burstein, Tetreault, Chodorow, Blanchard, & Andreyev, 2013).

“Holistic scoring” happens when an essay is reviewed by a human judge who assesses its quality and assigns a grade (Page & Petersen, 1995). Raters are instructed to read rapidly for a complete impression and consider all literary elements outlined in the rubric when using the holistic scoring technique (Burstein, 2003).

The holistic scoring method has been found to have certain advantages over the trait scoring method. This includes being more efficient in the scoring process and producing ratings that are more reliable (Koskey & Shermis, 2013; Page, Poggio, & Keith, 1997). However, one of the significant drawbacks of holistic rubrics is their inability to provide detailed diagnostic information about an examinee's performance. Since the criteria are combined and the feedback is not specific, the learners cannot understand the strengths and weaknesses of the essay and which characteristics of the writing influenced the rater's judgment (Koskey & Shermis, 2013).

When it comes to essay scores, they can be classified into one of four scales of measurement, which are nominal, ordinal, interval, or ratio. Rubrics that are commonly employed to evaluate essay responses on high-stakes exams are typically assessed on ordinal scales, which involve ranking the quality of a writing sample from lower to higher quality. These rubrics usually produce ratings within a range of 1 to 5 or 1 to 6 (Koskey & Shermis, 2013). The tests that include sections measuring writing ability, such as TOEFL, GMAT, or IELTS, provide similar scores in the writing section.

2.3. Human Scoring and Feedback

The teachers had been the sole source of writing development for many years in traditional writing classes for scoring and feedback. However, with the adoption of learner-centered approaches in writing education, coupled with advancements in technological tools, a shift in feedback practices occurred. Nowadays, feedback for improving writing skills is not limited to instructor-generated feedback but also includes peer feedback and automated feedback (AbuSeileek & Abualsha’r, 2014; Ravand & Rasekh, 2011)

Although human beings are at the core of writing evaluation, and even automated scoring systems have been developed based on the scores provided by them, it is difficult to comprehend what psychological elements constitute the way humans score essays. It is even more difficult to translate those elements into a computer program to create an automated scoring system (Page,1966).

2.3.1. Benefits of Human Scoring and Feedback

Students highly appreciate the feedback from teachers concerning their content and organization, as it is a human response to their essays and not generated by an AWE tool designed to identify errors through pattern recognition. Additionally, the research by Zhang and Hyland (2018) demonstrates that skilled ESL teachers can provide more thorough corrective feedback on the essays of the learners, addressing issues like number usage and abbreviations that are typically beyond the purview of AWE tools.

Human feedback is specific, pays attention to detail, and can discern what actions are necessary for the students to enhance their writing (Chen & Cheng, 2008; Huang, 2014).

Instructor feedback can encourage the improvement of the writing construct in the areas such as writing complexity and fluency (Link et al., 2020) by providing detailed feedback on content, development, coherence, and cohesion (Huang, 2014). As a result, the logical structure of arguments can best be evaluated by humans (Bridgeman, Trapani & Attali, 2012; Gao, 2021).

The feedback related to higher-order skills is provided better by the instructors; however, as research suggests, the errors related to lower-level skills are also better identified by humans on a greater number of error categories (Chen, Chiu, & Liao, 2009; Dikli, 2010; Dikli & Bleyle, 2014; Link et al., 2020; Zhang & Hyland, 2018). Instructors are better at identifying "hard-to-read text/proofread/garbled, connect sentences/coordinating conjunctions, comma splices, and run-on sentences" (Almusharraf & Alotaibi, 2022, p. 9) as well as errors related to collocation and syntax (Gao, 2021).

Finally, in the study by Huang (2014), the students emphasized the interaction they get from the teacher and the importance of negotiation, which is an essential part of the writing process.

2.3.2. Drawbacks of Human Scoring and Feedback

Ensuring consistency in scoring standards across various tasks and forms poses a significant challenge in human scoring of constructed response assessments as the interpretation of scoring criteria varies significantly among raters (Attali, 2013). Human-rater essay scoring frameworks have often faced issues related to subjectivity

and inconsistency in scoring. However, the implementation of Automated Essay Scoring (AES) can address these concerns by providing timely and accurate score predictions (Shin & Gierl, 2021).

The following example related to the inconsistency among human raters can be given. A group of 73 teachers from a school district in the United States, who had been trained in the same assessment method, graded a single paper. The scores given to the student essay varied from 50 to 96 on a 100-point scale, with a standard deviation of 9.6. Regarding letter grades, the paper was awarded an A by ten teachers, a B by 18, a C by 30, a D by nine, and an F by six teachers (Brimi, 2011).

According to studies, different elements can affect human writing evaluation, and human raters use diverse ways and strategies when making judgments (Wind & Engelhard, 2013). One problem that results in inconsistency among the raters is reader bias/drift. Rating errors or reader bias /drift can be described as "inadequacies of one sort or another in the ratings" (Saal et al., 1980, p.413). The outcomes of reader bias can be categorized into four. Firstly, the severity/leniency bias refers to the inclination of a rater to assign scores that are either too low (severity) or too high (leniency) in comparison to the appropriate level of performance demonstrated in the response. Secondly, the halo effect can be described as the rater's inability to differentiate between distinct and possibly unrelated aspects of a ratee's behavior. In addition, central tendency bias happens when evaluations are concentrated around the center of the rating scale. Finally, restriction of range, akin to central tendency bias, involves ratings that are clustered together, but in this case, the clustering can transpire at any point on the rating scale (Saal et al., 1980). The introduction of systematic reader errors or biases into the evaluation process results in variance unrelated to the measured construct, leading to a reduction in the validity of the scores (Messick, 1995).

Rater drift, which leads to biased scoring, is believed to stem from various factors related to the reader's characteristics, psychology, and rating conditions. Reader characteristics include experience in teaching, rating (Attali, 2016; Barkaoui, 2010b), content being assessed, and their linguistic and cultural backgrounds (Barkaoui, 2010a; Rao & Li, 2017). Reader psychology refers to internal factors such as fatigue, not remembering scoring procedures after breaks, attitude towards the prompt, and inability to disregard factors that do not impact the score, such as illegible handwriting. The rating conditions can also have an impact on the evaluation process, such as the

number of papers to be assessed, time pressure, and the desire to achieve high agreement rates (Braun, 1988; Hoskins & Wilson, 2001; Quellmalz, 1980) and also the mood in the room and the chat among the raters (Engelhard, 1994). Identifying rater drift is a challenging task, and therefore, the integration of automated scoring as a monitoring tool in the scoring process was expected to decrease the occurrence of central tendency drift (Lottridge et al., 2013).

Moreover, some academics (Alshakhi, 2019; Cotos, 2015) hold the opinion that assessing written texts is an arduous activity that entails numerous difficulties, such as devoting time to the evaluation process, giving personalized and timely feedback, and answering students' questions about feedback and the scoring guides. Besides, due to the number of students in a class and the teachers' busy schedules, teacher evaluation frequently takes the form of a summative score and is given many weeks after students submit their papers (Huang & Renandya, 2020).

Continuous training and multiple grading of a paper can increase interrater reliability and score consistency (Douglas, 2009). However, research indicates that the subjectivity element cannot be eliminated entirely (Wiseman, 2012). Therefore, automated scoring is an alternative to human scoring since it is an effective tool for minimizing scoring inaccuracies (Lottridge et al., 2013).

The concerns related to inconsistencies in human scoring are also encountered in teacher feedback. More attention is given to scoring because the stakes are high, and exam results determine milestones in students' lives. However, feedback is equally important and may be equally inconsistent. Providing feedback has been an essential aspect of enhancing the writing skills of students in their second language (Burston, 2001); however, consistently providing such feedback can pose challenges, particularly when dealing with a large number of students, leading to fewer writing assignments given in class (Burstein et al., 2003 as cited in Dikli, 2006).

Teacher feedback is not always consistent because of the errors human beings are likely to make. Therefore, it is difficult for the teachers to identify each error each time it occurs, and this situation creates confusion on the part of students as the message given by the teacher is not consistent (Ranalli et al., 2017). Moreover, providing accurate and consistent holistic feedback can become time-consuming, instructor-centered, and problematic. It leads to a process that places a heavy burden on teachers

who must correct a lot of written work completed by their students (Chapelle et al., 2015; Warschauer & Grimes, 2008; Wilson & Czik, 2016). As a result, using AWE tools as a formative assessment to assist writing instruction is perceived favorably when the issues related to human feedback are taken into account.

2.4. History of Automated Writing Scoring & Automated Writing Evaluation

The idea of grading essays using a computer was originally put forward by Ellis Page in 1966 when he asserted the immediate need to get help from computers as English teachers were overwhelmed by their workload. The first systems or tools created were for the purpose of scoring, and they were developed to be used in high-stakes exams together with human evaluation as summative assessment (Stevenson, 2016). While the validity and reliability of these AES tools are still being questioned, and many studies are being conducted to prove their validity and reliability (see Shermis & Burstein, 2003, 2013), a different aspect of automated scoring tools was discovered: evaluation. Then, they started to be called automated writing evaluation (AWE) systems. The word “evaluation” is critical here as it integrates feedback. This aspect led many researchers to include AWE tools as part of writing instruction and observe their impact.

Being the first person to put forward the idea of automated essay scoring (AES), Page (1966) suggests that as the first step in achieving success in AES, the AES model they would create should imitate or simulate expert raters in writing. However, the idea that AES should replace humans creates problems validating AES systems. One of the challenges is that if AES is intended to replace human evaluation, then it is crucial to demonstrate through AES validation that machine-generated scores assess the same skills or traits as human ratings. Still, it is almost impossible to know what constructs the process of human essay evaluation (Attali, 2013). A machine, in this case, a computer, cannot genuinely "read" a text and evaluate the grammar in the same manner as an experienced English instructor since it does not have a human brain (Attali, 2013; Elliot & Klobucar, 2013; Page & Petersen, 1995). Although the human rating process is not entirely understood, and it is difficult for a machine to emulate the same process, there are still several criteria that AWE systems rate a text’s quality accordingly, which are in Table 1.

Table 1. Different Dimensions of Essay Quality

Dimension	Description
Grammaticality	Grammar
Usage	Use of prepositions, word usage
Mechanics	Spelling, punctuation, capitalization
Style	Word choice, sentence structure variety
Relevance	Relevance of the content to the prompt
Organization	How well the essay is structured
Development	Development of ideas with examples
Cohesion	Appropriate use of transition phrases
Coherence	Appropriate transitions between ideas
Thesis Clarity	Clarity of the thesis
Persuasiveness	Convincingness of the major argument

Source: Ke & Ng, 2019, p. 6300

Based on these criteria and the systems and models it was developed accordingly, the e-rater projects a score that resembles a person. The computer does not interpret the essays better than the human raters; it just extrapolates the data we currently have more effectively (Elliot & Klobucar, 2013).

All automated essay scoring (AES) systems utilize different techniques, such as Bayesian text classification, statistical modeling, latent semantic analysis (LSA), natural language processing (NLP), and artificial intelligence (AI) (Shermis & Burstein, 2003). However, to determine essay scores, these techniques count on extracting linguistic and semantic features from a text (Allen, Jacovina & McNamara, 2016). In addition to this traditional machine learning-based AES that involves hand-engineered features, a deep-neural AES algorithm has been created. The term "hand-engineered features" typically denotes basic and intricate language features, which may reflect linguistic and non-linguistic aspects of writing quality (Shin & Gierl, 2021). On the other hand, prior research has shown that employing deep language features as part of a deep-neural algorithm can be a powerful strategy for achieving precise predictions in AES. Moreover, one crucial advantage of using deep language features is that they enable the simultaneous acquisition of pattern recognition skills and feature extraction capabilities through parallel learning (Latifi, 2016). Although

research on automated scoring using modern deep-neural algorithms yields promising results, only traditional AES models will be handled in this study.

2.4.1. Traditional AES/AWE Models

All the traditional AES/AWE models in this section count on extracting linguistic and semantic features from a text. The features used in these models and the way they assess the writing construct are a part of the writing theory (Attali, 2013). Automated essay scoring systems might effectively measure several characteristics of the writing construct directly. These characteristics, in other words, features include tense and aspect markers, place and time adverbials, pronouns, questions, nominalizations, passive forms, and subordination (Attali, 2013), as well as features related to fluency, lexical and syntactic complexity, and accuracy in writing (Wolfe-Quintero, Inagaki, & Kim, 1998), and world knowledge, text easibility (Shin & Gierl, 2021). The features that constitute ETS's e-rater measure grammar, usage, mechanics, vocabulary, and word length (Attali & Powers, 2009). Interpreting some features is obvious. A feature that highlights subject-verb agreement mistakes assesses grammar, word count in an essay measures, fluency, and the selection of words in relation to their developmental acquisition norms evaluates vocabulary level. On the other hand, it is feasible to create features based on the statistical correlation between a large number of little signals and the results of human essay scoring (Attali, 2013).

While establishing features, justification for the feature and its empirical relationship to other features and human ratings should be taken into account. Factor analysis can be implemented to present the correlations between features, and these correlation patterns offer additional support for their interpretation as indicators of certain aspects of writing quality (Attali, 2013). The initial step in automated scoring can be considered feature identification, and the features should be aggregated to provide essay scores in the following step. Aggregation or feature weighting can be accomplished by calculating a weighted average of the features. The weights in this calculation reflect the relative significance of the various features, and a variety of schemes are used for the aggregation of features. These schemes lead to scores with different meanings as different traits are underlined (Attali, 2013).

Overall, AES systems evaluate essays by applying predetermined features and using models to recognize patterns to produce precise grades (Zhang, 2013).

2.4.1.1. Trins and Proxes by PEG

> *A trin is the intrinsic variable of real interest to us. For example, we may be interested in a student's "aptness of word choice," or "diction." A prox, on the other hand, is some variable which it is hoped will approximate the variable of true interest. For example, the student with better diction will probably be the student who uses a less common vocabulary* (Page, 1966, p.240).

As suggested by Page and Petersen (1995), the distinction between human evaluation and an automated one was described via two terms: *trins* and *proxes*. While *trins* refer to intrinsic features of interest, *proxes* are approximations of these features. As Page (1994) puts it, they established a line between the factors we could measure and the variables we intended to measure. For instance, there was a correlation between the word count (prox) and the trin of fluency. If one wants to measure diction (trin), one needs to check word length (prox). Prepositions, relative pronouns, and other elements of speech served as proxes to the trin: the complexity of sentence construction (Page, 1994).

In their first trial - the first use of the PEG program -the results were stunning, with a correlation of .50, as Page puts it (1966). The variables or proxes included in this study are the proportion of uncommon words, the proportion of prepositions, the length of the essay, and the frequency of commas.

2.4.1.2. Latent Semantic Analysis (LSA)

So as to determine and evaluate the content of a text, AWE systems exploit a method named latent semantic analysis. LSA is a statistical modeling method that models word usage using massive corpora. After analyzing a sizable language sample being read by a machine, LSA shows the vocabulary used in the sample and any group of these words—such as a sentence, paragraph, or essay—as points in a very high-dimensional "semantic space." LSA produces statistics on word usage in a document, which is used to compare the content with reference documents on the topic. This is how AWE systems decide if the essay is semantically similar to the given prompt (Landauer, Foltz & Laham, 1998). However, Latent Semantic Analysis focuses only on the essay's conceptual content—the information it conveys—rather than syntax, style, or mechanics (Landauer, Laham & Foltz, 2003).

As Landauer et al. (1998) mention, the meaning representations extracted through Latent Semantic Analysis (LSA) show remarkable similarity to some human cognitive processes such as "developmental acquisition of recognition vocabulary, word-categorization, sentence-word semantic priming, discourse comprehension, and judgments of essay quality" (Landauer et al., 1998, p. 260).

LSA basically works with the presumption that each word in a text contributes to the development of its overall meaning and that a text's meaning is inherent in the words that make it up. However, two paragraphs with utterly different wording could have a similar meaning. To sum up, as the sum of the meanings of each word equals the meaning of the whole text, Latent Semantic Analysis works on that (Landauer et al., 2003).

There are some limitations to Latent Semantic Analysis. First, it does not employ morphology, syntactic relations, logic, or word order. However, surprisingly, it does excellent work extracting accurate impressions of passage and word meanings without these tools. Still, it is sometimes regarded as incomplete or likely to include errors (Landauer et al., 1998).

2.4.1.3. Natural Language Processing (NLP)

Natural language processing emerged in the 1950s as a combination of artificial intelligence and linguistics (Nadkarni, Ohno-Machado, & Chapman, 2011). NLP can be described as "*a range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of task applications*" (Liddy, 2001, p.1)

NLP-based tools execute a variety of language analyses making use of statistical and rule-based methods. While statistical methods are based on considerable corpora to model linguistic patterns, in rule-based methods, some rules are created to identify the structures of the language. Two types of modeling approaches to these methods are supervised and unsupervised. While supervised modeling approaches need data annotated by humans, unsupervised modeling does not require any annotations (Burstein, Tetreault, & Madnani, 2013).

NLP is a huge part of our daily lives as it is used for automated grammatical error detection in many software, Internet search engines, and machine translation (Shermis

et al., 2013). As part of NLP, machine translation employs computational analyses to do the translations between two languages. Automated summarization, which elicits the most related parts from a text, is another NLP tool that can be used in everyday life. Internet search engines, which need to evaluate the content typed in the search engine, have been developed based on NLP methods (Burstein et al., 2013a).

2.5. Different AES/AWE Systems

This section will provide brief information related to the most widely used commercial AES/AWE systems in the market. Table 2 shows the basic information about different tools.

Table 2. Different AES/AWE Systems

AES System	Developer	Author	Platform	Main Focus	Essay Scoring Mechanism	Features
PEG **Project Essay Grade**	Measurement Incorporated (MI)	Ellis Page	Mi Write erblearn .org	Style	Statistical	Trins & Proxes
IEA **Intelligent Essay Assessor**	Pearson Knowledge Technologies (PKT) Pearson Test of English (PTE)	Peter Foltz, Darrell Laham, Thomas K. Landauer	WriteTo Learn, Writing Coach, Summary Street	Content	Latent Semantic Analysis (LSA)	Content, Style, Mechanics
e-rater	ETS	Jill Burstein	Criterion	Content & Style	Regression Analysis, Natural Language Processing	Errors in Grammar, Errors in Word Usage, Errors in Mechanics, Style, Organization, Vocabulary
IntelliMetric	Vantage Learning	Scott Elliot	My Access!	Content & Style	Artificial Intelligence,	Focus and Unity,

					Natural Language Processing, Statistics	Development and Elaboration, Organization and Structure, Sentence Structure, Mechanics and Conventions (Elliot, 2003, p: 72)

Source: Adapted from Lim, Bong, Wong, & Lee, 2021, p: 1879

2.5.1. Project Essay Grade (PEG)

Automated essay scoring engine PEG, which belongs to Measurement Incorporated, grades student essays. To create a cutting-edge automated scoring engine, MI concentrated on combining the most recent developments in natural language processing, semantic and syntactic analysis, and classification techniques (Measurement Incorporated, 2020).

PEG evaluates essays based on writing quality (Page, 1966, 1994). The theory behind PEG includes the terms *trins* and *proxes* – as mentioned above in detail. PEG's objective is to predict scores that several skilled human judges would assign to a collection of related essays. In general, PEG employs statistical regression analysis to score the given essays. So far, PEG has created more than 500 trins to be utilized in essay scoring, such as fluency, diction, syntax, and structure (Measurement Incorporated, 2020). The ultimate goal is to minimize the distance between these trins and proxes (Page, 2003). Examples of specific proxes that may be assessed by PEG into a grade include average word length, the number of semicolons, and how rare a word is (Rudner & Gagne, 2000).

PEG's strong point lies in its ability to give scores that match those given by humans. In addition, the program can digitally monitor the errors in writing committed by its users. Nonetheless, PEG has received negative feedback for disregarding the meaning of essays and prioritizing surface structures instead (Chung & O'Neil, 1997). Due to its inability to recognize content-based elements such as organization and style, the system fails to give informative feedback to students (Dikli, 2006).

2.5.2. Intelligent Essay Assessor (IEA)

A combination of software tools called the Intelligent Essay Assessor (IEA), developed by Pearson Knowledge Technologies, is used to rate the quality of writing content. The IEA employs latent semantic analysis (LSA), a technique for determining the semantic similarity of words and passages from the text, and a computational model of how humans represent knowledge (Foltz et al., 1999)

How does Intelligent Essay Assessor work? The first method is called Holistic, which compares an ungraded essay to a group of prescored essays that yield a typical response quality range (Landauer et al., 2003). In various experiments, the Holistic Method's reliability coefficients were very close to inter-rater reliability coefficients (Foltz et al., 1999; Landauer et al., 2003). The second method employed by IES is comparing writings to a "gold standard" essay (Wolfe et al., 1998). In this scenario, a teacher can create their perfect essay, after which all student submissions would be evaluated based on how much they match the instructor's essay. The last method includes matching specific sentences from a student's essay against a collection of predetermined subtopics or textbook chapters. This makes it possible to evaluate if an essay sufficiently addresses certain subtopics (Foltz et al., 1999).

In contrast to other AES systems, IEA places greater emphasis on features related to content rather than features related to form. Nevertheless, this does not imply that IEA does not offer feedback on form-related elements such as grammar and punctuation in a text. While the system primarily relies on an LSA-based method to assess the excellence of an essay's content, it also gives grades and feedback on aspects like grammar, style, and mechanics (Landauer et al., 2003).

2.5.3. IntelliMetric and MY Access!

IntelliMetric is an advanced scoring program imitating the scoring procedure performed by human raters. It is based on the principles of cognitive processing, computational linguistics, and classification. Before use, the tool requires training with an array of previously rated essays that contain "known score" papers for each band in the scoring guide. These papers serve as the foundation for the program to deduce the scoring guide and the collective assessments made by the human scorers (Elliot, 2003). By utilizing AI and NLP technologies, IntelliMetric assesses more than 300 features related to semantics, syntax, and discourse in an essay. The features are categorized

into five overarching categories: Focus and unity, development and elaboration, organization and structure, sentence structure, mechanics, and conventions (Elliot, 2003).

MY Access! is an online writing evaluation system based on the IntelliMetric automated essay scoring system. The primary objective of the software is to provide learners with a writing platform that can quickly assess and offer constructive feedback, enabling them to make revisions to their essays (Vantage Learning, n.d., as cited in Dikli, 2006).

2.5.4. E-rater and Criterion

Natural Language Processing is the inherent computational methodology in the Electronic Essay Rater (e-rater) of the Educational Testing Service (ETS). E-rater is an advanced "Hybrid Feature Technology" which makes use of syntactic variation, discourse structure (similar to PEG), and content analysis (similar to IEA), and it is built to recognize textual features that mirror the writing characteristics listed in the scoring criteria for human readers (Rudner & Gagne, 2000; Burstein, 2003). E-rater contains three NLP-based modules to determine the features that are found in the scoring guide in the categories of syntax, discourse, and topic, and the features that are equal to these categories are as follows: syntactic variety, organization of ideas, and vocabulary use, respectively (Burstein, 2003).

The first module is E-rater's syntactic module. It involves a syntactic parser that identifies parts of speech and creates syntactic trees to figure out the syntactic variety of the text. The discourse module, on the other hand, labels cue words required for argument development and the organization of the ideas. Lastly, to identify the topic and evaluate vocabulary usage, the vector-space model, which is used to retrieve information, is employed as part of the E-rater (Burstein, 2003).

The discourse module was developed by Attali and Burstein (2006), who focused on creating a metric that evaluates the quality of discourse coherence. Evaluation of discourse coherence quality is essential because instructions given to human evaluators of essays in standardized writing exams require them to consider coherence, which is an essential factor in scoring. Burstein et al. (2013a) devised a set of annotations that closely matched the high-level descriptions outlined in the rubric criteria to construct a system. The aim was to consider linguistic information that could

have a positive impact on coherence quality in the writing of test-takers. This linguistic information included how well the main ideas in an essay are connected and how clear those connections are to the reader.

In addition, Beigman-Klebanov, Burstein, Madnani, Faulkner, and Tetreault (2012) added a system for sentiment analysis that can identify the sentiment and polarity (positive, negative, or neutral) present in essays written by students and test-takers. Sentiment analysis includes identifying personal opinions and opinions attributed to others and recognizing positive and negative statements. The research on sentiment analysis is pertinent to the assessment of argumentation in writing, whose importance has been emphasized by the Common Core State Standards Initiative (CCSSI). (Burstein et al., 2013b).

E-rater uses feature extraction to build a model and score the essay. The features related to grammar, word usage, mechanics, discourse elements for an essay, style, topical word usage, differential word use, content, the sophistication of essay words and register, and correct usage of prepositions and collocations (Burstein et al., 2013) are determined. A numerical value is assigned to each feature. Then, a linear regression equation is used to give appropriate weight to each feature according to the training sample. The new, unscored essays are scored by going through the same process and getting a grade compared to the training sample (Attali & Burstein, 2006).

The features related to grammar error detection, essay-based discourse analysis, content vector analysis, and detecting collocations use NLP approaches (Burstein et al., 2013). Developing features for the e-rater is closely aligned with the criteria outlined in the scoring rubric. This is done to guarantee that the relevant aspects of the writing construct are taken into consideration (Burstein, Beigman-Klebanov, Madnani, & Faulkner, 2013).

E-rater has been integrated into Criterion, a web-based instructional system through which students can practice writing and get feedback (Burstein, Chodorow, & Leacock, 2004; Li et al., 2015). Criterion comprises a set of software programs that can detect grammatical, usage, and mechanical errors and recognize components of discourse and style that may be undesirable in an essay. Along with providing immediate holistic scoring, the Criterion system offers personalized feedback based

on the specific assessments teachers typically use when evaluating student writing (Burstein et al., 2004).

Criterion incorporates different essay types: persuasive, descriptive, narrative, expository, cause and effect, comparison and contrast, process, argumentative, and issue through many levels starting from Grade 4 to college level, TOEFL and GRE included. The users can benefit from Writer's Handbook to understand their mistakes. In addition, there are spell check and thesaurus tools that aid students in their essay composition process. Criterion is a platform that allows students to be autonomous in their writing, but it also gives instructors a chance to step in and give their feedback on student essays if they wish.

2.6. Automated Scoring and Feedback

Automated Essay Scoring (AES) has brought numerous advantages, including increased consistency in scoring and finding mistakes, cost- and time-effective scoring (Attali, 2013), and the ability to offer immediate feedback to learners related to their performance (Latifi & Gierl, 2021).

Using an AES system ensures cost-effectiveness and reliability by preventing changes in the measured construct, variations in attitude towards different subgroups, and vulnerability to significant scoring errors by making sure that writing conventions are adhered to (Attali, 2013; Attali, Bridgeman, & Trapani, 2010).

Despite their apparent benefits, AES systems have some drawbacks. The first concern is that machines cannot "read," understand and interpret an essay as well as humans do (Ericsson & Haswell, 2006; Page, 2003; Warschauer & Ware, 2006), and as a result, it cannot truly evaluate the underlying writing construct like the quality of the arguments and the logic (Attali, 2013).

As the machine lacks comprehension of the content (Chung & O'Neil, 1997), it is plausible that exam-takers may devise techniques to generate text that is not reliant on their writing skills and is illogical. However, it enables them to attain higher scores from the machine (Attali, 2013). The studies revealed that it is possible to trick AES systems to get a higher score by writing perfectly grammatical but irrelevant, even repetitive essays (Higgins, Burstein, & Attali, 2006; Powers et al., 2001). AES tools like e-rater can be trained to detect text that is irrelevant to the given topic, but they

might not be sensitive enough to identify an essay that is only slightly off-topic (Bridgeman et al., 2012).

For AES systems to evaluate student essays accurately, they must undergo training using a vast collection of pre-scored essay samples (Burstein, 2003; Elliot, 2003; Landauer et al., 2003). The AES systems are limited to scoring pre-scored prompts from their own libraries. While teachers can assign their own prompts, the computer is not equipped to score essays with unfamiliar prompts as it lacks the necessary training to evaluate them (Dikli, 2006).

Finally, practitioners who are hesitant to adopt AES often express doubts regarding the transparency of modeling, the interpretability of algorithms, and the accuracy of the scores given (Zaidi, 2016).

2.6.1. Validity and Reliability of Automated Essay Scoring

Using AES in large-scale assessments is supported because of the many benefits it brings (Bereiter, 2003). Because the stakes are high in large-scale assessments, the validity and reliability of the AES tools have been questioned since they were used for the first time.

Many studies have been conducted to illustrate that AES systems are as reliable and as valid as human scoring, and some of the previous research showed a significant correlation between the AES and human raters (Attali, 2013; Dikli & Bleyle, 2014; Shermis, 2014). More studies on the reliability of AES tools have consistently found that the correlation between grades generated by AES systems and grades given by human scorers falls within the same range as the correlation between two human raters' scores (Warschauer & Ware, 2006). For example, in the TOEFL iBT independent task, the correlation between automated grades and human grades (.72) was higher than the correlation between two human scores (.68) (Bridgeman, 2013).

Studies related to the validity of PEG (Page et al., 1997; Shermis, Koch, Page, Keith & Harrington, 2002) and studies related to the validity of e-rater (Burstein, Kukich, Wolff, Lu & Chodorow, 1998) and studies related to the validity of IntelliMetric (Vantage Learning, 2001) yielded similar correlations as the two human raters.

Shermis (2014) notes that AES systems have been improved to such an extent that they can simulate the scores given by and agreed on by two raters in a consistent manner. They are objective and reliable, with easy scoring (Attali, 2013).

However, there are some studies reporting that AES tools are not reliable enough as the scores from an AES tool (MyAccess) and humans correlate moderately (Hoang & Kunnan, 2016), and AES tool (Criterion) do not grade the essay as reliably as human raters (McCurry, 2010). Moreover, in Wang and Brown's study (2007), the comparison of mean scores between IntelliMetric and human raters showed significant differences between the two groups. IntelliMetric assigned a significantly higher mean score compared to the mean score given by human raters.

2.6.2. Automated Feedback

Developing automated essay scoring systems to be used in large-scale assessments and trying to prove their validity and reliability is a highly investigated topic. However, it is crucial to devise an efficient approach that can also give personalized feedback together with analysis of a vast number of essays. Instructional-based AES systems try to attain this and aim to support the process of evaluating writing in classrooms. These systems are created to assist teachers rather than replace them (Dikli, 2006).

In addition to the summative assessment role of AWE systems, some AWE systems, such as MY Access! and Criterion, can be used as formative assessments in the classroom. These AWE systems provide feedback to the learners and allow them to produce multiple drafts depending on the feedback they get from the AWE system. The more AWE systems are used as formative assessments, the more credible they will become (Shermis & Burstein, 2003).

Providing feedback through AWE offers clear benefits compared to feedback given by teachers, particularly in terms of speed, ease of access, the ability to revise multiple times, and even the possibility of increasing learner autonomy (Chen & Cheng, 2008; Dikli, 2006). In addition, AWE tools are more reliable and objective (Rezaei & Lovorn, 2010), and AWE feedback is efficient and consistent (Gao, 2021).

On the other hand, AES systems are not exempt from criticism. Position statements on the assessment of writing by the Conference on College Composition and Communication (2004, 2009) oppose automated scoring. They assert that writing for a machine diminishes the validity of the evaluation since it conveys the message that

human communication is not valued. Additionally, because they are unable to determine the criteria used by the computer to evaluate the writing, they are also unable to identify any potential biases.

As they have originally been developed to score a high number of papers in high-stakes exams, they have been criticized for the reliability of the scores and the worries about the implications of writing for an audience that is not human. This criticism is about the social aspect of writing. As humans are social creatures and writing is a social construct, using AES systems to evaluate writing dehumanizes the writing construct (CCCC, 2009).

2.6.3. Human Scoring and Feedback vs. Automated Scoring and Feedback

AWE systems have been developed due to some problems encountered related to human scoring and feedback. They both have their own advantages and disadvantages.

The argument goes that whereas AES focuses on surface grammatical characteristics of an essay and semantically important vocabulary and rates the essays accordingly, human raters typically look at other aspects of writing besides grammar and vocabulary, such as content, development, coherence, and cohesion (Huang, 2014). Even though AWE is trusted with its feedback on lower-level features, human feedback on lower-level features is still more than the AWE tool with better quality-more accurate and focused (Chen et al., 2009; Dikli, 2010; Dikli & Bleyle, 2014; Link et al., 2020). As Dikli and Bleyle (2014) and Park (2019) found in their studies, Criterion and Grammarly identified fewer errors than human raters. In contrast, in Almusharraf & Alotaibi's (2022) study, Grammarly detected more errors than human raters. However, all these identified errors were related to style, word choice, format, and punctuation and were repeated throughout the texts. While these categories were the most prevalent errors detected by Grammarly, the error categories identified by human raters included "hard-to-read text/proofread/garbled, connect sentences/coordinating conjunctions, comma splices, and run-on sentences" (p. 9). On the other hand, Tömen (2022) illustrates that the feedback provided by Criterion on grammar and mechanics is similar to teacher feedback, and the teacher and Criterion have identified a similar quantity of errors.

In the study conducted by Zhang and Hyland (2018), they found that the feedback provided by teachers covered a more significant number of error categories compared

to the feedback generated by AWE tools, and the AWE system did not identify certain error categories, including articles and prepositions. However, while the teacher directly corrected all the student errors – 86.5 % of the time – the AWE feedback gave indirect feedback by highlighting the errors. No agreement has been reached in the academic literature about the efficacy of indirect feedback in promoting learning. However, this method can be helpful in increasing students' recognition of their errors (Ferris, 2011).

A study that employs Pigai – an AWE system developed by a Chinese developer to specifically address Chinese EFL learners (Jingxin & Razali, 2020) - discovered that AWE feedback could correctly identify spelling and most lexical mistakes. Still, it was less effective at detecting collocation and syntactic mistakes, which are more complicated since they include the ability to combine single words into phrases and sentences. Additionally, Pigai was unable to detect any logic issues, which were discovered to be prevalent in the student essays. The mistakes related to collocation, syntax, and logic were discovered by the human raters (Gao, 2021).

It is essential to indicate that AES systems make it easier to correct essay-type questions as they lessen the workload of instructors who would otherwise have to correct them manually. They minimize the potential for personal bias during the grading process. Moreover, more extended essays are valuable to human raters, and it has been found that e-rater is inclined to value long essays more than human raters (Bridgeman, 2013).

Today, using AES systems in the assessment of writing is on the rise, and AES systems are commonly employed in writing classes and as co-raters in the assessment of large-scale standardized writing, such as GMAT essays, where ETS has been using e-rater in conjunction with a human rater since 1999. This is done while still retaining the involvement of human raters (Dikli, 2006). Therefore, instead of choosing either the teacher feedback or the feedback from an AWE tool, combined feedback may be exploited in writing classrooms so that instructors can adapt their feedback approach and dedicate additional time to instructional or formative assessment to enhance the rhetorical structure of learner essays by utilizing AWE feedback (Zhang & Hyland, 2018).

2.7. The Impact of AWE in Classroom

Improving students' writing skills is arduous for teachers and students (Dikli & Bleyle, 2014). While teachers are supposed to give feedback on the drafts that students write for improvement, the students need to practice a lot for their own improvement. Because of limited resources, the maximum amount of practice cannot be done by both parties. Teachers face difficulties allocating sufficient time during the day to provide adequate feedback on students' work. For this reason, automated essay evaluation tools have come to our aid. These tools can be used as part of writing instruction to produce more writing and improve the quality of the lesson.

AWE feedback can potentially improve writing skills; however, there is currently insufficient research to prove its efficacy in L2 writing courses (Link et al., 2020).

2.7.1. Positive Effects of AWE Systems

2.7.1.1. Improvement in Writing Skills

AES provides students with feedback for every draft they write in no time, along with feedback on errors and relevant explanations (Kern & Warschauer, 2000; Milton, 2006; Ware & Warschauer, 2006). Providing immediate and constant corrective feedback as part of writing instruction encourages students to write more and enhance their writing accuracy (Hartshorn et al., 2010; Li et al., 2015).

Although the quality of the AWE feedback is being questioned, the students still prefer the AWE tools to be able to get immediate feedback (Grami, 2020). It is an advantage of AWE tools that they provide immediate holistic and analytical feedback (Dikli, 2006; Yeh, Liou, & Yu, 2007), and many students expressed their satisfaction related to the AWE tool because of the immediate feedback.

Students appear enthusiastic about obtaining prompt feedback on their writing and actively participate in editing while the assignment is still fresh in their minds. The ability to edit an essay several times at their own speed is also a benefit of receiving feedback through AWE (Warschauer & Ware, 2006). In the study by Yeh et al. (2007), college EFL students found the feedback provided by MY Access! helpful in revising their writing, particularly with regard to improving focus and meaning.

AWE tools provide the students with as many revisions as they like. Typically, students would engage in the revision of an assignment about four times when using the system, which is more revision practice than what would usually happen in a traditional classroom where teachers are grading the work. The outcome of utilizing the system demonstrated that, on average, student writing enhanced by one point on a six-point scale across those revisions (Foltz, Lochbaum, & Rosenstein, 2011 as cited in Foltz, Streeter, Lochbaum, & Landauer, 2013). Wang et al.'s (2013) study found similar results. The students who used Criterion over the semester got higher scores from the AWE tool and the human rater than those who did not practice with Criterion.

Many previous studies that employ AWE tools as part of writing instruction discovered that AWE tools improve students' lower-level writing skills, like grammar, vocabulary, spelling, and punctuation (e. g. Barrot, 2021; Ghufron & Rosyida, 2018; Jayavalan & Razali, 2018; Parra & Calero, 2019). In their study, Wang et al. (2013) explored the effects of an AWE tool on writing accuracy. The results of the post-tests showed that the students in the experimental group were significantly more accurate in capitalization, punctuation, the use of articles, tenses, different vocabulary, etc., than the control group. Moreover, when the pre-tests and post-tests of the learners in the experimental group were compared, it was obvious that the students who utilized an AWE tool made fewer mistakes in their post-test. The same results were taken in Liao's (2016) study too. In addition to that, the students displayed a positive attitude in terms of the AWE tool helping them improve their writing accuracy (Dikli & Bleyle, 2014; Lv, 2018), or they generally liked the feedback they got from the AWE tool (Huang & Renandya, 2020).

Another research by Li et al. (2015) reports that the ESL students' future revisions of the same paper might see a reduction in error rates in eight of the nine error categories (Word Choice, Verb Form, Articles, Pronoun, Run-on Sentence, Fragment, Sentence Structure, and Subject-Verb Agreement) thanks to feedback they got from Criterion. The learners appear to have been successful in converting the short-term gain from automated feedback into benefits in the long run, only in the Run-on Sentence error category; however, there was no statistically significant decline in any of the remaining eight error categories. There are other studies proving the positive impact of AWE tools in the long run, which means that students who use AWE tools to practice writing retain accuracy in their grammar (Lee, 2020).

It is worth noting that Grammarly provided the students with the ability to notice, which plays a significant role in writing development. As a result of the metalinguistic explanation provided by Grammarly, students could notice the errors they made (Barrot, 2021).

The impact of automated feedback on correcting grammatical, usage, mechanical, and stylistic errors among students was observed. Still, the degree and manner of this influence varied among different learners. Although AWE programs may assist students in rectifying grammar and mechanics errors (e.g., spelling and capitalization), they typically do not make any changes to their writing, like changes to collocations or sentence structure, even though they submit a few drafts (Ebyary & Windeatt, 2010).

The studies given above assert clearly that AWE evaluates a text's surface-level features, such as the number of words, the length of the sentences, and the range of the vocabulary. Still, it does not evaluate the rhetorical and stylistic devices used in the text (Huang, 2014).

2.7.1.2. Students' Attitude

The studies above indicated that students who used an AWE tool displayed a positive attitude towards these tools because they liked the idea of immediate feedback, multiple revisions, and improvement in their writing skills. Using AWE tools affects students' attitudes in many ways.

AWE feedback can be a valuable form of formative assessment, as it provides regular and process-focused diagnostic assessment during the revision, highlighting the active role of learners. Consequently, it can promote learner engagement (Zhang & Hyland, 2018)

AWE systems lead to more motivation in writing classes (Grimes & Warschauer, 2010; Warschauer & Grimes, 2008). They encourage students to be more motivated to write and produce more and longer content (Li et al., 2015; Li, Link, Ma, Yang, & Hegelheimer, 2014; Wang, 2013). Although this situation can be attributed to the novelty effect (Chen & Cheng, 2008), it motivates students to some extent.

Using an AWE tool promotes learner autonomy with the resources it provides and the functions it offers (Dikli, 2006; Ghufron & Rosyida, 2018; Warschauer & Grimes,

2008). Moreover, elevated learner autonomy helps students to become ready to produce more and better writing (Wang et al., 2013). Another study discovered that using an AWE tool as part of writing instruction improved the students' self-efficacy, and their scores on state tests have hence improved indirectly (Wilson & Roscoe, 2020). For example, Criterion allows for both teacher and peer feedback on the same platform, which leads to interactivity and students benefiting from different perspectives in addition to the feedback provided by the AWE tool.

2.7.1.3. Instructors' Attitude

Teachers prefer AWE tools in the hope that it would lessen their work (Dikli & Bleyle, 2014) and they do diminish the teachers' workload in providing feedback (Bai & Hu, 2017; Chen & Cheng, 2008; Warschauer & Grimes, 2008; Wilson & Czik, 2016). In Link et al.'s (2020) study, they found out that while the traditional teacher-only feedback included 22.60 feedback units per 1000 words in the control group, the teacher feedback supported by an AWE tool contained 10.15 feedback units, which means the teacher gave almost half as much feedback when the AWE was part of the writing instruction. In line with Link et al.'s (2020) study, teachers in Wilson and Czik's (2016) study report that using AWE tools saves them some time while giving feedback as the feedback on lower-level writing skills is handled by the AWE tool.

In the interviews conducted with the writing instructors in Li et al.'s (2015) study, the instructors believed in the positive effect of AWE corrective feedback in improving grammar and mechanics. Even though some of the instructors raised concerns related to the quality of the feedback, the research results illustrated that the AWE tool seemed to help students to write more accurately.

Although some studies report that the scores from an AWE tool are not reliable (Li et al., 2015), some instructors find AWE scores helpful in identifying who is at risk (Wilson, Olinghouse, McCoach, Santangelo, & Andrada, 2016) and some instructors use the scores from AWE tools as a forewarning, a benchmark for students before submitting their writing, and as part of classroom assessment (Li et al., 2014). Grimes and Warschauer's (2010) study reported that the AWE tool - MY Access! - provided help to the teachers related to classroom management.

2.7.2. Criticism towards AWE Systems

One of the most significant concerns related to AEE is the possibility of inappropriately rewarding or punishing particular aspects of examinees' writing, which poses a danger to its validity (Powers et al., 2001). In their 2001 study, Powers et al. wanted to see if the AWE tool, e-rater, could be fooled. The results of the study indicated that some essays written by writing experts were able to deceive the AEE tool. E-rater could not identify the problems such as repetition of a couple of paragraphs, repeating the same paragraph by slightly changing the topic sentence, and using only the features AWE awards and assigning the top score to these essays. In contrast, the grades of human raters assigned to these essays ranged from 1 to 3.

The questions that arise in the face of AWE are usually related to the fact that AWE cannot assess the true qualities of writing as well as human raters do. The students can trick the algorithm into getting a good grade (Shermis et al., 2013).

The results from Chen & Cheng's (2008) study demonstrate that the AWE feedback was unable to address students' writing issues sufficiently and in a meaningful way and cannot get what the students truly meant as teachers did (Wang et al., 2013). For this reason, teacher feedback is also required (Attali, Lewis & Steier, 2013; Benali, 2021; Li et al., 2014; Wali & Huijser, 2018). Some students asserted that having AWE feedback without complementary teacher feedback affected their attitude towards the AWE tool and the writing class negatively, and the students in the combined feedback condition showed the most improvement (Mohsen & Alshahrani, 2019).

The limits of automated scoring were a source of worry for many students, who noted that AWE encourages length, ignores coherence and richness of the content, dissuades the employment of nontraditional essay writing techniques, and there is too much emphasis on the use of transition words (Chen & Cheng, 2008; Wang et al., 2013). Teachers also mention the same problem that AWE favors length; correcting minor grammatical mistakes can improve the total score (Li et al., 2015). Studies on AWE show that AWE systems have mostly ignored other essential aspects of writing progress, such as complexity and fluency, in favor of accuracy (Link et al., 2020).

AWE systems lack the depth of any genuine interaction that may occur in the actual world, and the students regard the AWE tools negatively in that score (Chen & Cheng, 2008; Wang et al., 2013). Writing entails the students interacting with the reader and

being logical. However, as AES tools lack the necessary interaction, they fail to identify logical problems (Lv, 2018).

The feedback from an AWE tool can be unfocused, indirect, vague, formulaic, repetitive (Chen & Cheng, 2008; Wang et al., 2013), not specific, long (Dikli, 2010), non-dialogic and fallible (Bai & Hu, 2017). For example, Criterion may suggest revising a punctuation error by highlighting the whole sentence (Link et al., 2020), or it indicates that the sentence is wrong but does not provide further details related to the error, and this situation frustrates the students (Li et al., 2015). The inaccuracies or vagueness in the feedback by the AWE tool may confuse the students and lead them to delete or ignore the feedback or fail to correct it (Link et al., 2020; Ranalli, 2018; Wang et al., 2013). Additionally, Criterion failed to accurately identify the wrong word choices under the category of "Wrong or Missing Word," as it is highly challenging for a computer to discern the semantic linkages between words as well as humans can (Dikli & Bleyle, 2014). Because of this kind of feedback, research that has been conducted recently has revealed that applying AWE in L2 writing provides a considerable barrier for learners (Bai & Hu, 2017; Hockly, 2019; Mehrabi-Yazdi, 2018; Ranalli, 2018).

The feedback from the AWE tools is mostly provided in English, which is frustrating for the EFL students at lower levels as they have difficulty understanding the metalanguage and the feedback well (Wang et al., 2013). However, it is a fact that beginning or intermediate learners would benefit more from an AWE tool's assistance since it focuses primarily on mechanical precision and formal structure, which are seen as being more significant than meaning in early L2 writing acquisition (Chen & Cheng, 2008).

The primary issues related to the grammar feedback from the AWE tools were identified as the limited range of error types detected, and the inconsistency in the way errors were flagged (Chen & Cheng, 2006). The same situation occurs with Criterion feedback in which some of the identified error codes were correct, and some were not and most of the grammatical errors go unidentified (Han, Chodorow, & Leacock, 2006; Lavolette, Polio & Khang, 2015; Tetreault & Chodorow, 2008). Because of these, it has been found that AWE tools do not improve students' holistic writing quality (Huang & Renandya, 2020; Wilson & Roscoe, 2020)

Both the students and the instructors were not satisfied with the feedback they got from Criterion related to content and organization because, at present, AWE tools have limited capability to evaluate the intended meaning of a text (Chukharev-Hudilainen & Saricaoglu, 2016). For this reason, the instructors themselves preferred to give feedback on higher-order skills, like content development and organization, while leaving lower-level skills, like grammar and mechanics, to Criterion to deal with them. Some teachers criticized Criterion for encouraging students to follow a fixed writing pattern by acting as a template and preventing and even punishing students for writing creatively (Li et al., 2015).

In addition, the students using My Access! were not given any practical advice to aid in revising their writings. This was especially clear in the parts of coherence and content development. Furthermore, once an essay was flagged "off-topic," the students got no feedback and found this situation frustrating (Chen & Cheng, 2008).

In identifying organizational issues at the discourse level and off-topic content, the feedback from the AWE tool was insufficient. This revealed that the diagnostic efficacy of AWE feedback is still constrained and cannot be used as an independent instructional tool (Hoang & Kunnan, 2016). The feedback provided by AWE does not result in significant improvements in students' revision skills (Stevenson & Phakiti, 2014).

While some of the students who used an AWE tool as part of their writing classes did not find it helpful and did not trust the scores provided by the AWE tools, as there were discrepancies between the teacher score and the automated score (Chen & Cheng, 2008; Li et al., 2014), some students trusted it and some instructors still wanted to benefit from it (Grimes & Warschauer, 2010; Li et al., 2014).

Notably, the participants in Zhang's study (2017) were incentivized by the possibility of achieving higher overall scores through multiple revisions. They experienced a sense of discouragement when their scores did not improve despite repeated revisions.

2.8. Combined Teacher and AWE Feedback

Many studies (Chen & Cheng, 2008; Choi, 2010; Link et al., 2020;) preferred to use AWE tools together with the teacher feedback, not alone. However, one teacher in Chen & Cheng's (2008) study depended on AWE feedback alone, but the students

expressed their discontent at the end of the semester, mentioning that they needed teacher feedback to guide them. The AWE tool was mainly trusted to give feedback on lower-level skills, while teachers dealt with higher-order skills. (Link et al., 2020).

It must be indicated that in Wilson & Czik's study (2016), the students who got feedback from their teachers and an AWE tool improved their writing persistence. However, there was no significant difference in the final-draft writing quality.

In another study, the students who got teacher feedback improved in 8 out of 9 CAF (complexity, accuracy, fluency) measures, while those who used an AWE tool improved only in accuracy. However, it is essential to note that AWE students kept improving accuracy until the delayed post-test; the same situation was not observed with the teacher feedback group (Link et al., 2020). Lee (2020) also discovered that using an AWE tool improves writing fluency and complexity.

It is interesting to note that contrary to what is suggested, Wali and Huijser (2018) only used feedback from an AWE tool. Over half of the students reported that the AWE tool improved their grammar, content, and vocabulary. However, they still needed teacher feedback as well. In his study Choi (2010) found out that the students who used Criterion together with teacher feedback outperformed those who did not use Criterion or used Criterion optionally in terms of holistic quality and accuracy in grammar (Choi, 2010).

3. METHODOLOGY

The purpose of this study was to examine if using an AWE tool – Criterion – and a combination of automated and teacher feedback affect the writing skills and attitudes of EFL learners towards writing in English when compared to the students who obtained feedback solely from the instructor. The study also investigated the students' views on using Criterion as part of the writing class and having automated and teacher feedback together.

3.1. Research Design

This study used a naturalistic classroom-based method to determine if integrating an AWE system into EFL writing classes had an impact on students' writing performance. A mixed-methods research design where qualitative and quantitative data were collected was adopted because combining qualitative and quantitative research methods offers a more comprehensive comprehension of the research problem compared to utilizing only one of them; in other words, it provides "the best of both worlds" (Plano Clark & Creswell, 2015, p.385).

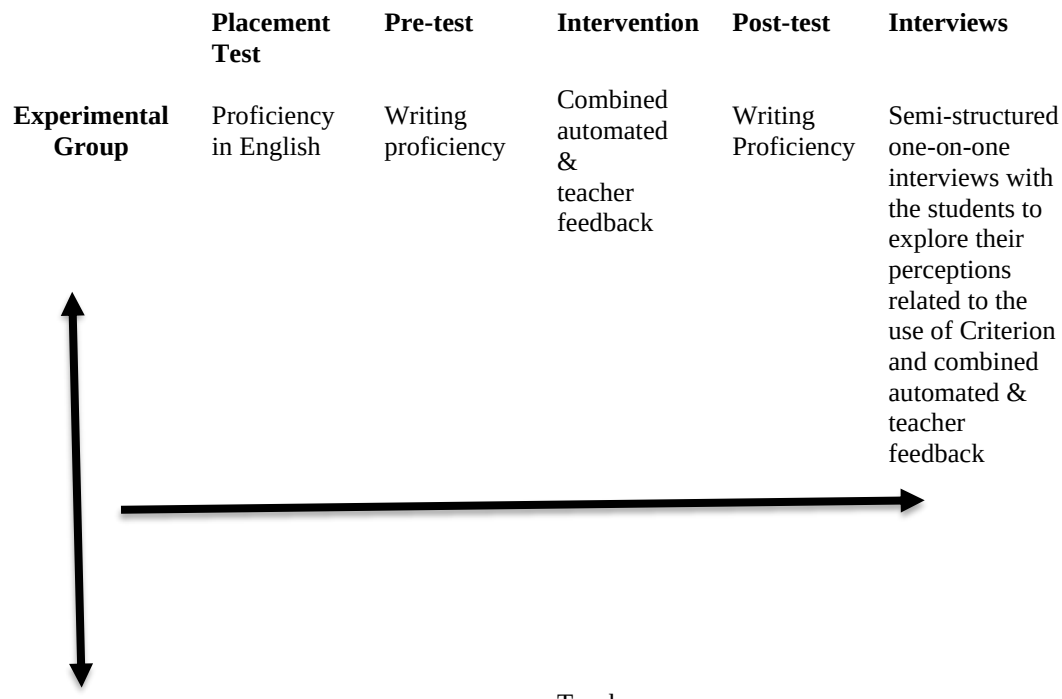
In the quantitative part of this study, a quasi-experimental research design which included an experimental group and a control group was utilized. The purpose was to investigate the effect of combined automated-teacher feedback in comparison to the condition where only full teacher feedback was provided. The reason why this research design was chosen is that experimental design is suitable for examining the effect of a treatment or intervention on a particular result (Creswell, 2009). True experimental research design would have been preferred as it is "the most rigorous test" of all the experimental research designs (Plano Clark & Creswell, 2015, p.195). However, implementing true experiment research is not possible in educational settings as the participants are already a part of a class and they cannot be assigned randomly (Cohen, Manion, Morrison, & Morrison, 2007; Creswell, 2012) and this led the researcher to have a quasi-experimental research design in which the experimental and control groups were formed by assigning two intact classes. Through this design and

quantitative research methods, the purpose was to investigate if the students' writing skills improved after receiving feedback from an automated tool and a teacher, as measured by pre- and post-tests.

When the researchers aim to investigate the perceptions and experiences of people regarding a specific concept, or a phenomenon, they collect data in the form of written or spoken expressions and utilize text analysis techniques, such as thematic analysis, to identify patterns and themes within the data (Plano Clark & Creswell, 2015). For this reason, the qualitative part of this study included data from the semi-structured interviews with the students. The data were thematically coded to understand the students' perspectives on the benefits and problems of using Criterion as part of a writing class and getting feedback.

It is important to note that collecting both quantitative and qualitative data was performed simultaneously, and the analysis of each type of data was conducted independently. Then the results were integrated to arrive at a comprehensive interpretation.

Figure 1. Research Design



As Figure 1 illustrates, during the intervention, the experimental group was given combined automated – teacher feedback, while the control group only obtained feedback from the instructor. After the treatment, both groups were given a post – test. The scores from both groups were then analyzed to determine the effect of Criterion use and the combined automated-teacher feedback on the overall grades of the learners. Moreover, one-on-one interviews were conducted with 18 students from the experimental group to examine the students' perceptions related to using an AWE tool and combined automated – teacher feedback.

3.2. The Setting of the Study

The study was administered in the spring semester of the 2022-2023 academic year in the Basic English Department of the School of Foreign Languages at a state university in İstanbul, Turkey. At the beginning of the fall semester, students willing to start their departments take an institutional English Proficiency Exam (EPE) to prove their level of English. If a student gets 60 or above out of 100, he/she is considered proficient and allowed to start their studies in their department. However, the students who do not take the exam or fail are supposed to enroll in the preparatory program and, as a result, take an institutional English Placement Test to be placed in the appropriate level. In the fall semester, there are three levels: A1, A2, and B1, which are designed according to the Common European Framework of Reference. The students are placed at a suitable level in accordance with the institutional English Placement Test results. As all the participants in this study studied at the A1 level in the fall semester, information related to this level will be provided here. In the fall semester, A1 level students got 24 hour-instruction per week incorporating four skills: reading, writing, listening, and speaking and grammar. Moreover, students in this level were supposed to be at the A2 level, ready to begin the B1 level, when the term ended. However, there is no institutional test to determine the level of the students at the end of the fall semester. As for the writing instruction, it was at the paragraph level, and students learned how to write an opinion paragraph and a cause-effect paragraph for five hours each week using the writing material prepared by the Curriculum and Material Development Office of the institution.

When this study was conducted in the spring semester, A1 level students started the B1 level program in different classrooms with different instructors. The B1 level curriculum covers all four skills: reading, writing, listening, and speaking and grammar in 20 hours per week. Within these 20 hours of instruction, four hours were allocated for the writing skill, four hours for the reading skill, three hours for listening and speaking skills, and the remaining nine hours for the main course and grammar. The writing curriculum included teaching an opinion essay, a cause-effect essay, and various activities supporting the writing process. The writing material prepared by the Curriculum and Material Development Office of the institution was used in the writing classes.

3.3. Participants

3.3.1. Students

The researcher, who was also the instructor of the writing classes, implemented the study with two different intact groups as part of their writing lessons. One class that benefited from the automated feedback was the experimental group, while the other class that got teacher feedback was the control group. The students in both classes were from different majors and enrolled in the preparatory program to be proficient in English before they started their departments. To be able to do that, these students need to pass the institutional English Proficiency Exam (EPE) with a score of 60 or above, and the writing section of the exam is worth 25 points. The writing task in the exam includes two options, and the students are supposed to write either an opinion /persuasive essay or a cause & effect essay on diverse topics.

There were 26 students on the roster of both classes at the beginning of the semester. However, the number of students who actively participated in the study, completed the pre- and post-tests and attempted all the assignments and drafts dropped to 22 in the experimental group and 16 in the control group, mainly because the classes were delivered online and the students experienced some problems related to absenteeism and motivation.

The participants were EFL learners whose L1 was Turkish. There were 11 male and 11 female learners in the experimental group, while there were ten male and six female learners in the control group. Their ages ranged from 18 to 21. Moreover, the students

were non-English majors, and Table 3 shows their departments. It is significant to indicate that the students in the experimental group have never used Criterion before.

Table 3. The Departments of the Students

Experimental Group	Control Group
S1. Biomedical Engineering	S1. Chemistry
S2. Mathematical Engineering	S2. Architecture
S3. Civil Engineering	S3. Ship Machinery Management Engineering
S4. Environmental Engineering	S4. Computer Engineering
S5. City and Region Planning	S5. Mechatronic Engineering
S6. Metallurgy and Materials Engineering	S6. Civil Engineering
S7. Metallurgy and Materials Engineering	S7. Mechanical Engineering
S8. City and Region Planning	S8. Ship Machinery Management Engineering
S9. Food Engineering	S9. Naval Architecture and Marine Engineering
S10. Environmental Engineering	S10. Statistics
S11. Architecture	S11. Computer Engineering
S12. Mechatronic Engineering	S12. Electrical Engineering
S13. Ship Machinery Management Engineering	S13. Environmental Engineering
S14. Bioengineering	S14. Mathematical Engineering
S15. Metallurgy and Materials Engineering	S15. Map Engineering
S16. Chemical Engineering	S16. Environmental Engineering
S17. Architecture	
S18. Chemical Engineering	
S19. Civil Engineering	
S20. Environmental Engineering	
S21. Chemistry	
S22. Architecture	

In order to ensure that both the experimental group and the control group had equal English proficiency levels, a placement test was given to all participants before the study commenced. Oxford Quick Placement Test included 60 questions, each worth one point, so the scores of the placement test were calculated out of 60 points. The descriptive statistics calculated using SPSS pertaining to the placement test results are provided in Table 4. The mean scores for each group showed that according to the categories included in the Oxford Quick Placement Test, the students had an A2 level of English and were ready to start getting instruction at the B1 level.

Table 4. Descriptive Statistics of the Placement Test

	N	Mean	SD	Min	Max
Experimental Group	22	26.5	6.262	14	41
Control Group	16	24.88	7.051	11	41

The next step in interpreting the placement test results was to see if the experimental and control groups differed in any way. The number of students in both groups was pretty low. Therefore, Mann-Whitney U Test, a non-parametric technique, was employed to see whether there was a difference between the two groups. Table 5 shows the results of the Mann-Whitney U Test. According to Pallant (2020), if the value of the Sig. column is above .05; the two groups have no significant difference. Therefore, the results indicate that the difference between the groups is not significant, which means the two groups were identical and had almost the same level of English proficiency. These two groups were similar in terms of their level of English; they were suitable for this study.

Table 5. Mann-Whitney U Test to Compare the Placement Test Results of the Experimental and the Control Groups

Total N	38
Mann-Whitney U	152.500
Wilcoxon W	288.500

Test Statistic	152.500
Standard Error	33.712
Standardized Test Statistic	-.697
Asymptotic Sig. (2-sided test)	.486
Exact Sig. (2-sided test)	.492

Participants in the study were asked for their informed consent and were promised anonymity while sharing their writing and opinions, as required by the institution's ethical committee.

3.3.2. The Instructor and the Raters

The researcher was the teacher of both classes and the first rater of all the pre- and post-tests. The researcher has the experience of almost 12 years in teaching and assessing writing. All the pre- and post-tests of both groups were assessed by the researcher and a second rater with similar experience. Four other raters assessed either the pre-tests or the post-tests as the second rater. The researcher and these raters were full-time employees at a state university's preparatory program.

As for the grading of the student essays, a holistic scoring approach was employed throughout the study. The grades given by Criterion are based on a holistic rubric created by ETS (Appendix 1), the same rubric that the AWE tool employs to grade the papers of the students who use Criterion. Also, this rubric is very similar to the TOEFL IBT Independent Writing Rubric, which only differs in the range of the scores. While the TOEFL rubric gives a score between 0 and 5, the Criterion rubric gives a score between 1 and 6. The experimental group was evaluated based on a holistic score, and the control group was also evaluated according to the same score. Moreover, pre-tests and post-tests were graded using the same rubric.

Prior to commencing the scoring procedure, the researcher and all other raters convened for a meeting to deliberate on the evaluation criteria outlined in the rubric and come to a consensus regarding the expectations. Then, the researcher and the raters evaluated the essays independently without any interaction or discussion about the scores they gave.

3.4. Data Collection Instruments

This study adopts a mixed-method research design. Therefore, quantitative and qualitative data collection tools were utilized during the research. The data collection instruments explained below include a placement test, pre- and post-tests, writing assignments, Criterion Rubric, the AWE system Criterion, Google Classroom, and Google Documents, and semi-structured one-on-one student interviews.

3.4.1. Placement test

Prior to the treatment, Oxford Quick Placement Test was given to the experimental and control groups through Google Classroom at the beginning of the semester to ensure that all the students had the same level of English. The test, whose reliability has been reported to be close to 0.90 (Geranpayeh, 2003; Jones, 2013), consists of 60 questions. When the placement test results were checked for this study, there was no significant difference between the experimental and the control groups. It means that they had similar levels of English proficiency.

3.4.2. Pre-test

A pre-test was given to the experimental and control groups through Google Classroom at the beginning of the spring semester to compare the results with the post-test so that the difference between the two groups could be observed. It was given as a timed task, which students had to complete in 75 minutes. The students were supposed to write an opinion/persuasive essay of 250 words on a topic provided by the instructor. The topic was chosen from the Criterion Topic Library, which consists of over 400 writing topics across various genres. These prompts are available for different proficiency levels, ranging from 4^{th} grade to university level, as well as TOEFL and GRE topics (Dikli, 2006; ETS, 2019). The students were supposed to practice opinion/persuasive essays in the first half of the spring semester according to the curriculum of the preparatory program. Therefore, the selected genre from the Criterion Topic Library was persuasive. The level preferred for this study was TOEFL because the wording of the topics was very similar to those in our institutional exams. The topics were determined according to this genre and level. However, because they were used for a TOEFL exam, some topics were too abstract and required a high level of proficiency to comprehend and produce ideas. For this reason, the topics were

selected according to the way they were worded, their concreteness, and considering students' English proficiency, interests, and familiarity. Topic familiarity is crucial in ensuring that all students have a fair opportunity to be successful (Weigle, 2012). Keeping all these in mind, the following topic was chosen by the researcher to set as the pre-test writing task:

> Do you agree or disagree with the following statement?
>
> *Only people who earn a lot of money are successful.*
>
> Use specific reasons and examples to support your answer.

The pre-tests of both groups were rated by the researcher and another rater independently using Criterion's Rubric. To determine the consistency between the scores the two raters gave, the Spearman Rank Order Correlation (rho) was utilized as a measure of inter-rater reliability. According to Pallant (2020), to be able to calculate inter-rater reliability using parametric tests, each group should have at least 20 cases. As the control group in this study incorporates 16 participants, the non-parametric alternative was chosen. As each test was evaluated by the researcher and a different rater, inter-rater reliability for each test and each group was calculated. For the experimental group, Spearman rho was .736, and it was .902 for the control group. These values indicate a strong correlation between the raters because, according to Dancey and Reidy (2007), the correlation between .07 and .09 is "strong," or according to Cohen (1988), the correlation is "large."

In writing assessment, the most significant aspect to be considered is the inconsistency of scoring that arises from the rating behaviors of the raters (Huang & Foote, 2010). Maintaining consistency among and within raters is crucial in writing assessment because it verifies the precision of grades given to student essays and is significantly associated with the impartiality of writing assessment (Weigle, 2002).

3.4.3. Post-test

After the intervention, a post-test was conducted to observe the improvement in students' writing and writing proficiency.

The conditions the post-test was conducted were almost the same as the pre-test. Because of a devastating earthquake that struck the southeastern part of Turkey on 6th February 2023, it was decided by the Higher Education Council that the classes would

be conducted online during the spring semester. For this reason, both tests were conducted online. Both the pre-test and post-test writing tasks used essay topics from the same level of English and genre to ensure the reliability of both tasks.

The two differences between the pre- and post-test are the platform on which they were submitted and the topics that were chosen. These differences emerge from the fact that the writing task of the midterm exam was used as the post-test. For this reason, students used the distance education system of the university to submit their writings. The topics were provided by the Testing Office of our institution as part of the exam, and they could not be changed.

There were two topics to be chosen in the post-test since it was the midterm exam in our institution. The following topics were given to the students:

1. *Social media plays an important role in learning a new language.* Do you agree or disagree? Write an Opinion Essay explaining your opinion. Write at least 250 words.

2. *Some people think that it is more comfortable to read e-books.* Do you agree or disagree? Write an Opinion Essay explaining your opinion. Write at least 250 words.

The post-tests of both groups were rated by the researcher and another rater independently using Criterion's Rubric. To determine the consistency between the scores given by the two raters, the Spearman Rank Order Correlation (rho) was utilized as a measure of inter-rater reliability. According to Pallant (2020), to be able to calculate inter-rater reliability using parametric tests, each group should have at least 20 cases. As the control group in this study incorporates 16 participants, the non-parametric alternative was chosen. As each test was evaluated by the researcher and a different rater, inter-rater reliability for each test and each group were calculated. For the experimental group, Spearman's rho was .841 and .724 for the control group. These values indicate a strong correlation between the raters because, according to Dancey and Reidy (2007), the correlation between .07 and .09 is "strong," or according to Cohen (1988), the correlation is "large."

To discover if there was a significant difference in the writing proficiency of both groups, thanks to the different feedback they received, the scores from pre- and post-tests were compared.

3.4.4. Weekly Writing Practices

The instruction in the writing class was given by the researcher following the same syllabus and provided to both the experimental group and control group students in the course of the intervention. The students in both groups were supposed to write five essays in the same genre, an opinion / persuasive essay. Both groups started writing their essays during lesson hours and submitted their drafts through an AWE tool – Criterion – before next week's lesson. The experimental group wrote as many drafts as they wanted, got feedback from Criterion, rewrote, and reached the final draft. On the other hand, the control group wrote a single draft through Google Documents on Google Classroom during this time. Both groups received teacher-written feedback before next week's class on these drafts and oral feedback from the teacher during lesson hours. The students were allowed to use any resources that would aid them in English and topic development.

3.4.5. Essay Evaluation Rubric

A holistic scoring rubric consisting of 6 bands of proficiency (Appendix 1) was used, with each level representing a score on a 6-point scale. While score 1 refers to the lowest proficiency, score 6 refers to the highest writing proficiency. The raters were prohibited from using half-points.

3.4.6. One-on-one Interviews

Upon the completion of the treatment and the post-test, 18 students from the experimental group volunteered to participate in semi-structured one-on-one interviews voluntarily. The purpose of these interviews was to gather comprehensive information about the students' perceptions of using an AWE tool and receiving feedback from a combination of an automated tool and a teacher, as well as the reasons for their perceptions. The interview questions were developed by the researcher as a result of reviewing the literature and consulting with her advisor to ensure their appropriateness. These questions were designed to explore the participants' experiences and perceptions of the subject. Seven open-ended questions (see Appendix 2) were created to gather information from participants about their experiences and perceptions of the writing and feedback they got. The purpose of these questions was to explore the students' viewpoints in detail.

The researcher conducted the interviews via Zoom, a communication platform that permits its users to communicate via video or audio. Before the interviews, the students were briefed on the study's objective and the interview process. They were additionally notified that their personal information would remain undisclosed, and the interview would be recorded exclusively for research purposes. Upon receiving their consent, the researcher conducted the interviews in Turkish, the students' native language, as it was deemed that they would be more comfortable expressing themselves in it. Each interview lasted approximately 15 minutes and was recorded and transcribed verbatim.

3.5. Data Collection Procedure

The primary purpose of this study was to answer research questions. To be able to serve this purpose, data were collected throughout the spring semester of the 2022 – 2023 academic term in the way that will be mentioned in this section. The AWE tool employed during this study was Criterion, whose vendor is ETS. The researcher first contacted Criterion’s global team via e-mail and asked for a complimentary trial account for research purposes. Necessary permissions were taken from the Basic English Department of the School of Foreign Languages and Ethics Committee of the university (see Appendix 3). The researcher told the administrative staff about having two different writing classes for the spring semester because, usually, only one writing class is given to each instructor. Two intact B1 level classes were chosen for the study. While one class was assigned to the experimental condition, the other was assigned to the control condition randomly.

The experimental and control groups' students were briefed on the study's objectives and procedures and given written informed consent forms (see Appendices 4 and 5) to ensure they understood their rights and any potential risks the study might pose. The researcher met with the experimental group for extra 1-hour training in addition to that week’s contact hours to acquaint them with the Criterion platform. As each student was granted an individual account in Criterion, they were instructed on how to set up a student account, access the assignment provided by the instructor, create an outline, work on the writing task, submit a draft, and make edits, as well as how to check teacher comments. After the students created their own student accounts, they were trained on how to receive automated and teacher feedback on the platform and, most importantly, how to interpret this feedback using a model essay written by the

researcher. Moreover, the researcher provided the students with information on how to use online resources, including the Writer's Handbook, together with a summary of the handbook summarized and interpreted by the researcher and the other materials that could be found on the Criterion platform. On the other hand, the control group was trained to use Google Classroom and Google Documents during the contact hours in the second week because this was how they would submit their essays and get feedback. After the classroom on Google Classroom was created, the students were invited to enroll. Then, the researcher explained how to interpret the error codes (Appendix 6), including the written teacher feedback, and how to correct them on a model essay. The training was not conducted in the first week of lessons because those hours were allocated for teaching how to write an opinion / persuasive essay.

The curriculum for the writing course included teaching two essay types: opinion /persuasive essay and cause & effect essay for the spring semester. Within this study, the students in both groups concentrated on learning how to write an opinion /persuasive essay and practicing it for six weeks. While the experimental group used the AWE tool Criterion to practice and get feedback, the control group did their writing through Google Documents on Google Classroom and got teacher feedback.

One four-hour session each week for six weeks was dedicated to this study. The first thing that needed to be done in the first week was to conduct a placement test to see if the students in both groups were genuinely at the same English proficiency level. The test results indicated that both classes were at the same level, which is A2. This level means that the students completed the A2 level and were ready to get an education at the B1 level. To reduce the impact of factors that have the potential to influence the outcomes, such as the curriculum, teaching methods, the topics of the essays, and writing procedures were identical for both sets of participants, and they were instructed by the same teacher. Throughout the intervention, both groups of students were provided with training in writing that followed the process approach, which involves a recurrent process that includes planning, drafting, and revising.

The students in both groups were taught how to write an opinion/persuasive essay in week 1. After completing the training, the students submitted a single-draft pre-test without any revisions. The pre-tests of both groups were graded by two human raters: one is the researcher herself, and the other is another instructor from the same institution using Criterion’s holistic scoring guide. A total of five opinion/persuasive

essays whose topics were chosen from the Criterion Topic Library by the researcher were written by all the students.

3.5.1. Intervention Procedure Each Week

The whole study took seven weeks: While the first week was allocated for the pre-test and the training on how to write an opinion/persuasive essay, the final week (week 7) was for the post-test. In weeks 2, 3, 4, 5, and 6, the students practiced writing opinion/persuasive essays on different platforms and got feedback from different resources.

3.5.1.1. Intervention and Feedback Procedure for the Control Group

Starting from week 2, the students in the control group wrote opinion/persuasive essays on the topics provided by the researcher. The same topics as the experimental group were given to the control group to control the variables as much as possible.

After the training on how to use Google Classroom and Google Documents and how to interpret the error correction codes and teacher feedback, the students enrolled in the class created for them on Google Classroom. After that, the instructor/the researcher created their first writing practice and assigned it to every student in the classroom. Then, the control group started to write their first essay; 75 minutes were allocated to students before they submitted their first drafts. Because it took time for the instructor to evaluate those essays, the written feedback was provided by the instructor outside of class hours through Google Classroom. The students were asked to review the written teacher feedback before they came to the next writing class the following week.

In week 3, the instructor started the lesson by giving oral feedback to each student. The instructor invited each student to a breakout room on Zoom and gave their one-on-one feedback. Each feedback session took around 5 minutes, in which the students asked about the parts of the feedback they did not understand, and the instructor mentioned the most prominent mistakes that the students made for that specific draft and gave advice on how to correct those mistakes and improve their writing. Following the teacher's feedback, the students in the control group wrote their second drafts and got written teacher feedback for the second draft too. After the students were done with their second drafts, the next topic for the following writing practice was provided

through Google Classroom by the instructor. They were given 75 minutes to complete the first draft of the new writing topic. This way, the 4-hour contact session ended, and these activities continued in the following weeks in this order until week 7, and a total of five essays with two drafts were written by the students in the control group.

Besides the procedure during the writing classes, the details of the teacher feedback should be included in this section. The instructor gave written and oral feedback to the first drafts and only written feedback to the second drafts. The instructor focused on the form, including grammar, usage, and mechanics in her feedback. The instructor highlighted the errors related to the words, phrases, or sentences that required modification or enhancement indirectly. The institutional error correction codes (see Appendix 6) were employed by the instructor to indicate the type of mistake made by the students. The error correction codes aided the instructor in providing indirect feedback to identify the problematic areas in the essays of the learners without providing the correct structures. At the start of the intervention during the training session, the teacher provided a list that included examples for each error and explained the meaning of each code to the students.

As these error correction codes solely pertain to the mechanics and grammar, the instructor also included notes recommending revisions regarding the content and organization. Specifically, these notes were related to the organization of the opinion/persuasive essay, the presence or absence of thesis statements, topic sentences, certain paragraphs, the reasoning within the body paragraphs and developing ideas, and the usage of different vocabulary—this technique combined form and content-focused approaches instead of using only one. The teacher spent around 15 minutes giving written feedback on each essay. The written feedback was followed by an oral feedback session in which the students asked questions related to the written feedback, and the instructor mentioned the most prominent errors and gave advice on improving their essay writing skills. Then, the students wrote the second draft after that. Another written teacher feedback was provided to the students on the second draft, focusing on whether they corrected their mistakes from the first draft and adding feedback related to polishing up the essay. During the intervention, the students in the control group wrote a total of 160 drafts.

3.5.1.2. Intervention and Feedback Procedure for the Experimental Group

Starting from week 2, the students in the experimental group wrote opinion/persuasive essays on the topics provided by the researcher through Criterion. To control the variables as much as possible, the same topics were given to both groups.

After the training on how to use Criterion and learning how to interpret the feedback from Criterion and the instructor, the students enrolled in the class that was created for them on Criterion. After that, the instructor/the researcher created their first writing practice and assigned it to every student in the classroom. Then, the experimental group started to write their first essay; 75 minutes were allocated to students before they submitted their first drafts. Because it takes no time for Criterion to evaluate those essays, following the submission of each draft, the experimental group instantly got automated feedback through Criterion. To make students utilize this feedback properly and encourage them to submit multiple drafts, they were asked to increase their overall score on Criterion. According to a study by Li et al. (2014), establishing an overall benchmark score that students must attain before handing in their essays to the instructor could enhance learner interaction with AWE feedback. This is because students are reportedly more driven and purposeful when utilizing AWE feedback and engaging in multiple revisions. As a result of the feedback they got and with the purpose of increasing the overall score, they wrote the second draft and got immediate feedback from Criterion one more time during class hours. As the students were allowed to write as many drafts as they wished to maximize their scores, they were given five days after the class to do that. When they were done with their drafts, they got written feedback from the instructor outside of the class hours through Criterion before the next week's class. The students were asked to review the written teacher feedback before they came to the next writing class the following week. The students in the experimental group had access to Criterion both inside and outside of class.

The structure of the writing classes was formed in the following way for the following weeks. The instructor started the lesson by giving oral feedback to each student, and the instructor invited each student to a breakout room on Zoom and gave their one-on-one feedback. Each feedback session took around 5 minutes in which the students asked about the parts of the feedback they did not understand both from the instructor and Criterion and the instructor mentioned the most prominent mistakes the students made for that specific draft and gave advice on how to correct those mistakes and

improve their writing. Combined feedback was chosen in this study because, in their study, Zhang and Hyland (2018) found that combining both types of feedback in classroom settings might be more beneficial. Thus, they suggested that future studies should investigate the integration of both teacher and AWE feedback within the same writing assignment. In other words, there should be a "division of labor" between human evaluators and AWE tools. While human evaluators would assess the global aspects of writing, AWE would concentrate on the local aspects of writing (Attali et al., 2013). Moreover, to bridge the social gap of negotiating meaning in writing, additional feedback from teachers or one-on-one oral feedback from the teachers can be exploited (Huang, 2014).

Following the oral teacher feedback, the students in the experimental group started writing the new assignment for that week. They were given 75 minutes to complete the first draft of the new writing topic. When they submitted their first drafts, they immediately got automated feedback and started working on their second drafts based on the feedback from Criterion. This way, the 4-hour contact session ended, and these activities continued in the following weeks in this order until week 7. A total of five essays with multiple drafts were written by the students in the experimental group.

Besides the procedure during the writing classes, the details of the teacher feedback should be included in this section. The instructor gave written and oral feedback to the final drafts besides the automated feedback the students got to the initial drafts. As Table 6 illustrates, the analysis of feedback based on Criterion traits encompasses five error categories found in student writing: grammar, usage, mechanics, style, and organization and development (Link et al., 2020). For this reason, the teacher feedback did not focus on errors related to grammar, usage, mechanics, and style because they were already handled by the AWE tool. However, there were some mistakes that were not identified by Criterion, or the students did not understand and correct them properly, so the teacher feedback had to focus on them. Moreover, most instructor feedback addressed content and organization because Criterion provided generic feedback in this area. Specifically, the feedback was related to the organization of the opinion/persuasive essay, the presence or absence of thesis statements, topic sentences, certain paragraphs, the reasoning within the body paragraphs and developing ideas, and the usage of various vocabulary. The teacher spent around 15 minutes giving written feedback on each essay. The written feedback was followed by an oral

feedback session in which the students asked their questions related to the automated feedback and written teacher feedback, and the instructor mentioned the most prominent errors and gave advice on the betterment of their essay writing skills. Then, the students wrote the other drafts after that. During the intervention, the students in the experimental group wrote around 350 drafts on Criterion.

Table 6. List of Categories and Subcategories of Criterion Feedback

Feedback categories	**Feedback subcategories**
Grammar	Fragments Run-on Sentences Garbled Sentences Subject-verb Agreement Ill-formed Verbs Pronoun Errors Possessive Errors Wrong or Missing Word
Usage	Determiner Noun Agreement Wrong Article Missing or Extra Article Confused Word Wrong Form of Word Faulty Comparisons Preposition Error Nonstandard Word Form Negation Error Wrong Part of Speech
Mechanics	Spelling Capitalize Proper Noun Missing Initial Capital Letter in a Sentence Missing Question Mark Missing Final Punctuation Missing Apostrophe Missing Comma Hyphen Error Fused Words Compound Words Duplicates Extra Comma
Style	Repetition of Words Inappropriate Words or Phrases Sentences Beginning with Coordinating Conjunction Short Sentences Long Sentences

	Passive Voice
Organization & Development	Introduction, Thesis, Main ideas, Supporting ideas, Transitional words and phrases, Conclusion

Source: Li et al., 2015, pp. 15-16 & Chapelle et al., 2015, p. 389

In week 7, the students in both groups took the post-test, which was graded in the same way as the pre-test. After the post-test, semi-structured interviews with 18 students from the experimental group were conducted. The interviews were conducted in the participants' primary language and lasted around 15 minutes each. They were recorded, transcribed verbatim, and then translated into English. The main idea was to determine whether it was possible to combine two flawed scoring methods effectively to enhance the overall precision of scoring (Lottridge et al., 2013). Table 7 illustrates the overall plan for the study.

Table 7. Study Plan

Date	Experimental Group	Control Group
Week 1 **27 Feb – 3 March**	Training on how to write an opinion/persuasive essay Training on Automated Scoring System (Criterion): how to register and how to interpret the feedback Pre-test	Training on how to write an opinion/persuasive essay Pre-test
Week 2 **6-10 March**	Writing Practice 1 – Draft 1 Automated feedback Writing Practice 1 – Draft 2 (Students were allowed to write as many drafts as they wished according to the feedback from Criterion)	Registering Google Classroom and training on Error Correction Codes and how to use Google Documents Writing Practice 1 – Draft 1
After the class	The teacher provided written feedback on the final draft of	The teacher provided written feedback on the 1st draft of *Writing*

	Writing Practice 1 through Criterion.	*Practice 1* using Google Documents.
Week 3 **13-17 March**	The teacher provided oral feedback on the final draft of *Writing Practice 1* in breakout rooms on Zoom. Writing Practice 2 – Draft 1 Automated feedback Writing Practice 2 – Draft 2 (Students were allowed to write as many drafts as they wished according to the feedback from Criterion)	The teacher provided oral feedback on the 1st draft of *Writing Practice 1* in breakout rooms on Zoom. Writing Practice 1 – Draft 2 Teacher written feedback Writing Practice 2 – Draft 1
After the class	The teacher provided written feedback on the final draft of *Writing Practice 2* through Criterion.	The teacher provided written feedback on the 1st draft of *Writing Practice 2* using Google Documents.
Week 4 **20-24 March**	The teacher provided oral feedback on the final draft of *Writing Practice 2* in breakout rooms on Zoom. Writing Practice 3 – Draft 1 Automated feedback Writing Practice 3 – Draft 2 (Students were allowed to write as many drafts as they wished according to the feedback from Criterion)	The teacher provided oral feedback on the 1st draft of *Writing Practice 2* in breakout rooms on Zoom. Writing Practice 2 – Draft 2 Teacher written feedback Writing Practice 3 – Draft 1
After the class	The teacher provided written feedback on the final draft of *Writing Practice 3* through Criterion.	The teacher provided written feedback on the 1st draft of *Writing Practice 3* using Google Documents.
Week 5 **27-31 March**	The teacher provided oral feedback on the final draft of *Writing Practice 3* in breakout rooms on Zoom. Writing Practice 4 – Draft 1 Automated feedback	The teacher provided oral feedback on the 1st draft of *Writing Practice 3* in breakout rooms on Zoom. Writing Practice 3 – Draft 2 Teacher written feedback

	Writing Practice 4 – Draft 2 (Students were allowed to write as many drafts as they wished according to the feedback from Criterion)	Writing Practice 4 – Draft 1
After the class	The teacher provided written feedback on the final draft of *Writing Practice 4* through Criterion.	The teacher provided written feedback on the 1st draft of *Writing Practice 4* using Google Documents.
Week 6 **3-7 April**	The teacher provided oral feedback on the final draft of *Writing Practice 4* in breakout rooms on Zoom.	The teacher provided oral feedback on the 1st draft of *Writing Practice 4* in breakout rooms on Zoom.
	Writing Practice 5 – Draft 1 Automated feedback	Writing Practice 4 – Draft 2 Teacher written feedback
	Writing Practice 5 – Draft 2 (Students were allowed to write as many drafts as they wished according to the feedback from Criterion)	Writing Practice 5 – Draft 1 Teacher written and oral feedback Writing Practice 5 – Draft 2 Teacher written feedback
Week 7 **10-14 April**	Post-test (Midterm Exam)	Post-test (Midterm Exam)

3.5.1.2.1. Diagnostic Feedback of Criterion / E-rater

The e-rater scoring system utilizes linguistically based characteristics of an essay and employs a statistical model that connects these characteristics with the overall writing quality of the essay. This system assigns a score to the essay on a scale of 1 (poor) to 6 (excellent) (Higgins et al., 2006). Besides, it offers detailed feedback in five areas: grammar, usage, mechanics, style, organization, and development.

The AWE tool employed in the present study is Criterion developed by ETS. Criterion was chosen for this study for a couple of reasons. First, it is one of the most widely used commercial tools available. Second, it has been used to rate TOEFL and GRE writing sections together with a human rater; the scores are determined by averaging both scores (ETS, 2019). Therefore, much research has been conducted on the validity and reliability of Criterion (e. g. Link et al., 2020; Shermis & Burstein, 2003, 2013). Lastly, Criterion rates and gives feedback on grammar, usage, mechanics, style,

organization and development and different essay types such as persuasive, descriptive, narrative, expository, cause and effect, comparison and contrast, process, argumentative, issue through many levels starting from Grade 4 to college level, TOEFL and GRE are included (Dikli, 2006; ETS, 2019).

Criterion generally offers indirect feedback by identifying errors (without making the corrections) and occasionally provides recommendations on the appropriate form. The error corrections provided are metalinguistic in nature and often come with a brief grammatical explanation. The feedback is not specific to any particular error and covers all the errors detected by Criterion. However, certain errors may not be detected by the system until others have been addressed (Lavolette et al., 2015).

Criterion offers three distinct users: Administrator, Instructor, and Student. In order to initiate an assignment using Criterion, the first step is creating a class and enrolling the students in that class. After that, teachers are allowed to add an assignment to the system for the specific class. Teachers can specify the level, genre, topic, deadline, time limit, and number of attempts during the assignment creation process. Instructors can choose from the Criterion topic library, which features over 400 essay topics ranging from 4th grade to college-level writing, TOEFL, and GRE, or they can create their own topic.

Prior to commencing their written response, students have the option to utilize the Criterion planning tool to generate an outline. The tool includes the standard outline format and templates tailored for different essay types. In this study, because the genre we focused on was opinion/persuasive essay, we only used a persuasive essay template. This tool aims to assist students during the prewriting phase of their writing process. Once the essay is handed in, Criterion promptly provides a holistic score between 1 and 6, together with feedback on various writing traits (see Figure 2). To be able to interpret these scores and trait-level feedback, the students could click on the buttons indicated in red squares in Figure 2. After receiving feedback from the program, students have the opportunity to modify their essays according to the provided feedback and subsequently hand in their revised work for further evaluation. The system limits students to a total of 10 attempts.

Figure 2. Criterion Holistic and Trait Level Scores

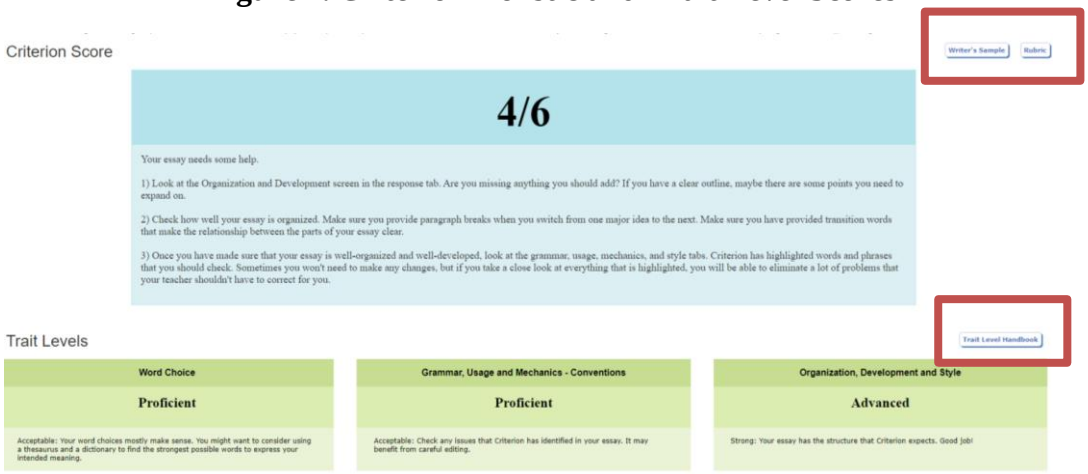
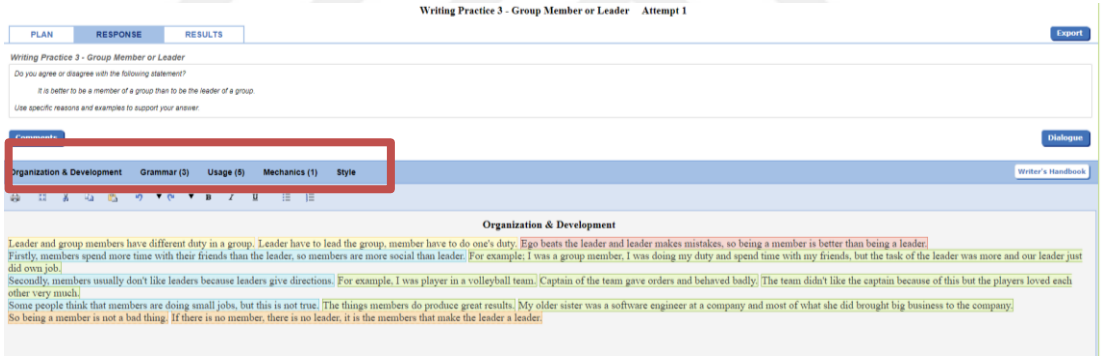


Figure 3 illustrates the color-coded feedback on Organization & Development, and the categories above (indicated in red squares) show the number of mistakes pertaining to each category and subcategory.

Figure 3. Criterion Feedback on Different Categories



When the students click on each category and subcategory, they can see the mistakes they made and get advice on how to correct these mistakes (see Figure 4).

Figure 4. Highlighted Errors and Advisory Messages



In addition to the automated feedback, the students got teacher feedback through Criterion, and they could reach the teacher feedback through the button indicated in the red square in Figure 5.

Figure 5. Teacher Feedback on Criterion



3.6. Data Analysis

With the purpose of answering the research questions, quantitative and qualitative data were collected within the span of this study. Various methods of analysis were utilized to examine the data collected from the participants.

3.6.1. Quantitative Data Analysis

The results of the pre- and post-tests of both experimental and control groups constitute quantitative data. The Statistical Package for Social Sciences (SPSS), version 26, was used to analyze the quantitative data acquired from placement test results and the scoring of pre-test and post-test writing tasks.

This program was utilized first to observe any differences between the experimental and control groups in their English level. The results of the placement test were entered on SPSS, and a codebook was created. As the number of participants in each group was relatively small, the Mann-Whitney U Test was conducted to compare the results of both groups.

Secondly, the program was used to check the inter-rater reliability between the researcher and the second rater who graded either pre- or post-tests of both groups. The Spearman Rank Order Correlation (rho) coefficient was calculated to get this result.

In the final analyses, via the SPSS program, the pre-test and post-test scores were analyzed both within and between groups to assess how the two types of feedback and

using Criterion affected the learners' writing performance. Because the Kolmogorov-Smirnov test and Histograms showed a non-normal distribution for the post-test results, and the number of participants for each group was relatively small, non-parametric tests were employed. Therefore, the Mann-Whitney U Test was run using the grades from the pre-tests to see if any differences between the experimental and control groups existed in their writing ability before the intervention. When the intervention was over and the post-tests were done, Wilcoxon Signed Rank Test was conducted to see if both groups improved thanks to the intervention from pre-test to post-test. After that, the Mann-Whitney U Test was conducted one more time on the post-test scores to observe if there was a significant difference between the two groups after the treatment they got. Finally, gain scores per student were calculated, and the Mann-Whitney U Test was used to see if there were any significant difference between the two groups in their gain scores.

3.6.2. Qualitative Data Analysis

The students' responses from one-on-one interviews, done after the study was over and the post-tests were finished, constitute the qualitative data in this study. Qualitative studies are commonly used in the social sciences and education and yield detailed and immersive depictions of human experience (Ary, Jacobs, Irvine, & Walker, 2009). By conducting one-on-one interviews and collecting qualitative data, this study aims to understand the participants' attitudes towards using an AWE tool as part of their writing class and their perceptions related to the efficiency of this tool and the combined automated-teacher feedback they got.

The qualitative data analysis spiral follows the following steps, as mentioned by Creswell (2013): organizing the data, reading and memoing, describing, classifying, and interpreting data into codes and themes, and interpreting the data and representing and visualizing the data. Qualitative analysis was conducted on the student responses by categorizing similar answers and displaying their frequency. Following Creswell's (2013) steps, the researcher began the qualitative analysis by recording and transcribing the interviews verbatim. After creating a file with the questions and all the transcribed student responses, the researcher reviewed the data multiple times, searching for similarities and differences. The researcher grouped similar data into categories and continually compared new data to previous data to ensure it was

appropriately categorized. Lastly, the researcher merged these categories into more general themes.

To verify the accuracy of the results, another coder who is a language teacher with a PhD degree specialized in qualitative data analysis was tasked with creating codes and themes based on the same data. There was a substantial agreement between the two coders regarding the coding and categorization of the data.

4. RESULTS

This chapter provides a comprehensive presentation of the data gathered through quantitative and qualitative data collection methods, all within the scope of the research questions, which aim to examine whether incorporating an AWE tool into writing instruction affects the writing performance and attitudes of Turkish EFL students when compared to traditional writing instruction with teacher feedback and to explore the students' perceptions of using an AWE tool in the writing class.

The results for the following research questions will be provided in this chapter:

RQ1. Do the use of Criterion and receiving combined automated-teacher feedback affect Turkish EFL students' writing performance significantly compared to teacher feedback only?

RQ2. What are the experimental group students' perceptions of combined-automated teacher feedback and their attitudes towards using Criterion as part of their writing classes?

Quantitative data was collected through pre-tests and post-tests, whereas qualitative data was gathered through one-on-one interviews. The quantitative and qualitative data analysis was done separately, and the presentation of the results was done in accordance with the two research questions given above. The quantitative data analysis, which addresses RQ1, was discussed first. After that, the results of the qualitative data analysis, which pertain to RQ2, are presented.

4.1. Quantitative Data Analysis Results

To address the first research question (RQ1) related to the effect of AWE-integrated writing instruction and combined automated-teacher feedback on Turkish EFL students' writing performance in comparison to traditional teacher-led instruction, the quantitative data from pre- and post-tests were analyzed using SPSS 26.0 for both descriptive and inferential statistics.

4.1.1. The Results of the Pre-test and Post-test

So as to provide data for the first research question, the pre- and post-test results of both the experimental and control groups were evaluated, and comparisons were made within and between the groups. Prior to applying any statistical analyses, the normality of the pre-test and post-test scores was checked to see if these scores showed normal distribution. Therefore, the type of analysis was determined. When the test of normality (Kolmogorov-Smirnov test) was done, the results of the post-test displayed a non-normal distribution. Also, the number of cases in both groups was relatively small. Therefore, non-parametric tests were conducted to see the differences between and within the groups. Table 8 displays the descriptive statistics for pre-tests for the experimental and control groups.

Table 8. Descriptive Statistics for Pre-tests of Both Groups

		N	M	SD	Min.	Max.
Pre-test	**Experimental**	22	3.46	1.09	1.5	5.25
	Control	16	3.93	1.48	1.5	6

The Mann-Whitney U Test was run using the scores from the pre-tests to observe if there were any differences between the experimental group and the control group in their writing ability before the intervention. Table 9 indicates the results of the Mann-Whitney U Test.

Table 9. Mann-Whitney U Test to Compare the Pre-test Scores of the Experimental and the Control Groups

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Pre-test Scores is the same across categories of Intervention Type.	Independent-Samples Mann-Whitney U Test	.312[a]	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .050.
a. Exact significance is displayed for this test.

Total N	38
Mann-Whitney U	211.000
Wilcoxon W	347.000
Test Statistic	211.000
Standard Error	33.539
Standardized Test Statistic	1.044
Asymptotic Sig.(2-sided test)	.297
Exact Sig.(2-sided test)	.312

As illustrated in Table 8, there are some differences in the mean scores of the experimental group (M=3.46) and the control group (M=3.93). However, the results of the Mann-Whitney U Test present no significant difference between the two groups regarding their writing performance before the intervention started (U=211.000, p=.312>.05).

When the intervention was over and the post-tests were finished, Wilcoxon Signed Rank Test was conducted to see if the experimental group improved thanks to Criterion as part of the writing class and combined automated-teacher feedback from pre-test to post-test. Table 10 illustrates the results of the Wilcoxon Signed Rank Test for the experimental group.

Table 10. Wilcoxon Signed Rank Test to Compare the Pre-test and the Post-test Scores of the Experimental Group

	Post-test Scores - Pre-test Scores
Z	-2.642[b]
Asymp. Sig. (2-tailed)	.008

b. Based on negative ranks.

When the value for Asymp. Sig. (2-tailed) is lower than .05, it means that the difference between the two scores is significant (Pallant, 2020). The results of the Wilcoxon Signed Rank Test for the experimental group show that there is a statistically significant difference between the pre-test and post-test results, meaning that the treatment in which the students used Criterion as part of their writing classes and got

combined automated-teacher feedback enhanced the writing performance of the students in the experimental group.

When the intervention was over, and the post-tests were done, Wilcoxon Signed Rank Test was run to see if the control group improved thanks to the intervention in which they practiced writing essays and got traditional teacher feedback from pre-test to post-test. Table 11 illustrates the results of the Wilcoxon Signed Rank Test for the control group.

Table 11. Wilcoxon Signed Rank Test to Compare the Pre-test and the Post-test Scores of the Control Group

	Post-test Scores - Pre-test Scores
Z	-1.315[b]
Asymp. Sig. (2-tailed)	.189

b. Based on negative ranks.

When the value for Asymp. Sig. (2-tailed) is lower than .05, it means that the difference between the two different scores is significant (Pallant, 2020), which means if the same value is higher than .05, the result is not statistically significant. The Wilcoxon Signed Rank Test results for the control group demonstrate that the difference between the pre-test and post-test results is not statistically significant. This means that the students who got traditional teacher feedback did not benefit from the treatment and did not improve their writing performance much.

Finally, the results of the post-tests were compared for both the experimental and the control groups. Table 12 displays the descriptive statistics for post-tests for both groups.

Table 12. Descriptive Statistics for Post-tests of Both Groups

		N	M	SD	Min.	Max.
Post-test	**Experimental**	22	4.25	0.97	1.5	5.5
	Control	16	4.43	0.62	3.5	5.5

The Mann-Whitney U Test was conducted using the post-test scores to observe any differences between the experimental group and the control group in their writing ability after the intervention. The aim was to see if the intervention helped one of the groups more. Table 13 shows the results of the Mann-Whitney U Test.

Table 13. Mann-Whitney U Test to Compare the Post-test Scores of the Experimental and the Control Groups

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Post-test Scores is the same across categories of Intervention Type.	Independent-Samples Mann-Whitney U Test	.759[a]	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .050.
a. Exact significance is displayed for this test.

Total N	38
Mann-Whitney U	186.500
Wilcoxon W	322.500
Test Statistic	186.500
Standard Error	33.177
Standardized Test Statistic	.316
Asymptotic Sig.(2-sided test)	.752
Exact Sig.(2-sided test)	.759

As illustrated in Table 12, the mean scores of the experimental group (M=4.25, SD=0.97) and the control group (M=4.43, SD=0.62) are close to each other in the post-test. Therefore, the results of the Mann-Whitney U Test present no significant difference between the two groups regarding their writing performance after the intervention was completed (U=186.500, p=.759>.05).

In addition to all these analyses, a final gain score analysis was conducted. Gain scores per student were calculated, and Table 14 provides the descriptive statistics for gain scores for the experimental and the control groups.

Table 14. Descriptive Statistics for Gain Scores of Both Groups

		N	M	SD
Gain Scores	**Experimental**	22	0.78	1.17
	Control	16	0.50	1.43

The Mann-Whitney U Test was used to see if there were any significant difference between the two groups in their gain scores. Table 15 shows the results of the Mann-Whitney U Test.

Table 15. Mann-Whitney U Test to Compare the Gain Scores of the Experimental and the Control Groups

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Gain Scores is the same across categories of Intervention Type.	Independent-Samples Mann-Whitney U Test	.569[a]	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .050.
a. Exact significance is displayed for this test.

Total N	38
Mann-Whitney U	156.500
Wilcoxon W	292.500
Test Statistic	156.500
Standard Error	33.548
Standardized Test Statistic	-.581
Asymptotic Sig.(2-sided test)	.561
Exact Sig.(2-sided test)	.569

The results of the Mann-Whitney U Test present no significant difference between the two groups regarding their gain scores after the intervention was completed (U=156.500, p=.569>.05).

The results of the quantitative analysis show that there was no difference between the students in both conditions in their final post-test scores. However, the students in the experimental group presented a significant improvement in their writing performance from the pre-test to the post-test. It can be concluded that using Criterion in the writing classes and getting combined automated–teacher feedback helped the experimental group improve more than the students in the control group.

4.2. Qualitative Data Analysis Results

In this part, the examination of the information acquired from one-on-one interviews conducted with the students in the experimental group is presented. With the purpose of discovering the thoughts and experiences of the students regarding the combined automated-teacher feedback and the attitude towards the use of Criterion as part of their writing class, one-on-one interviews were conducted at the end of the treatment. 18 of the 22 participants from the experimental group participated in one-on-one interviews. The information obtained from the one-on-one interviews was utilized to address the second research question (RQ2): What are the experimental group students' perceptions of combined-automated teacher feedback and their attitudes towards using Criterion as part of their writing classes? Below are the findings from the one-on-one interviews conducted.

4.2.1. The Results of the One-on-one Interviews

The researcher composed seven questions (Appendix 2) depending on the information from the literature and the observations throughout the study. The purpose of these questions was to guide the students to express their perceptions and attitude toward the process properly. The answers provided by the students were presented in the order the questions were asked during the interviews. The responses to the questions were coded, and themes were formed, as illustrated in the tables below. The tables also show the frequencies and percentages of the codes and themes.

Q1. What do you think about the scores given by Criterion? To what extent did they reflect your writing performance?

Table 16. Reflections about the Accuracy of the Scores Given by Criterion

Themes	Codes	*f*	*%*
Accurate Scoring	Accurate scoring in general	11	35.48%
	Accurate scoring when the topic is given from the topic library	3	9.67%
Inaccurate Scoring	Lower grades when the topic is given by the instructor	4	12.90%
	Higher grades in general	4	12.90%
Insensitivity Issues	Not sensitive to word count	1	3.22%
	Impact of basic mistakes on the final score	1	3.22%
	Insensitive to some mistakes	1	3.22%
	Rewarding longer sentences	1	3.22%
	Favoring certain expressions	1	3.22%
	Insensitive to content	2	6.45%
	Rewarding longer essays	2	6.45%
Total		31	100%

Table 16 provides the codes and themes that emerged from the students' responses to the first question. When the students were asked about the accuracy of the scores provided by Criterion between 1 to 6, 11 of them expressed that Criterion reflected their writing performance correctly through the scores it gave, as shown in the following statement:

"It usually scored my essay correctly, when I wrote good essays, I got a high score. However, for example, if I repeated the same vocabulary over and over again, Criterion gave me a lower score. It evaluates perfectly."

Three other students specifically expressed that the Criterion gave accurate scores only when the topic was chosen from the topic library, and it gave lower scores when the topic was chosen randomly by the instructor. In addition, four students expressed that they got higher scores than they deserved from Criterion. One more theme emerged from the students' responses: 'insensitivity.' The students gave some reasons why they thought Criterion scored incorrectly in the following statements:

"I was not sure if the program took the word limit into account. We were supposed to write a 250-word essay, but although I wrote 180 words per essay, I still got a high score."

"Criterion gave higher scores to my essays because I was writing longer than required. Once I realized that I did my best to write essays as long as possible to get a high score."

"I think the longer sentences you write, the higher grades Criterion gives to your essays."

Moreover, there are two students who mentioned Criterion's insensitivity towards the content, which led them to think that Criterion scores were inaccurate because it did not take content into consideration.

Q2. What did you do to increase the score that you got from Criterion?

Table 17. Students' Strategies to Increase their Scores on Criterion

Themes	**Codes**	***f***	***%***
Correcting the mistakes	Mistakes highlighted by Criterion	12	21.81%
	Grammar mistakes	11	20%
	Organizational mistakes	5	9.09%
	Vocabulary mistakes	7	12.72%
	Punctuation mistakes	5	9.09%
	Spelling mistakes	2	3.63%
Total		42	76.36%
Extra Efforts for Improving Writing Performance	Increasing word count	2	3.63%
	Rewriting some parts	1	1.81%
	Studying sentence structures and vocabulary	3	5.45%
	Doing research on the essay topic	1	1.81%
	Using a variety of tenses	1	1.81%
	Studying with a friend	1	1.81%
	Improving reading skills	1	1.81%
	Spending time on the later drafts	3	5.45%
Total		55	100%

Table 17 indicates the codes and themes that emerged from the responses the students gave to question 2 during the interviews. The students basically did two things to improve their score on Criterion. Most students (76.36%) expressed that they only corrected the mistakes highlighted by Criterion and increased their score. Correcting the mistakes highlighted by Criterion included the mistakes related to grammar, organization, vocabulary, punctuation, and spelling. Most students utilized one of

these strategies to improve their Criterion scores, as highlighted in the following statements:

"Criterion gave feedback mostly on grammar, vocabulary, and punctuation, and when I corrected all the highlighted mistakes, my score automatically increased."

"This is how I increased my score: The first drafts included many mistakes, especially mistakes related to grammar and inappropriate vocabulary use. I learned the correct grammar structures and vocabulary and corrected those mistakes. As a result, my score increased directly."

However, the rest of the students put extra effort into increasing their Criterion scores. They improved their writing skills in general in addition to correcting the mistakes highlighted by Criterion. These efforts included increasing word count, rewriting some parts, studying sentence structures and vocabulary, researching the essay topic to come up with better ideas and examples, using a variety of tenses, studying with a friend, improving reading skills, and spending more time on the later drafts thinking about the feedback they got. Some of these strategies can be seen in the following statements:

"If Criterion did not give much feedback on vocabulary, I tried to improve my vocabulary by trying to write different words, and I rewrote some parts that I did not like in the subsequent drafts."

"I researched the topic online when it was a bit different and difficult to write immediately."

"When I got feedback from Criterion, I thought about it for a while and rewrote the whole essay. When you spend time thinking about your mistakes and the missing parts, you write much better essays in the following drafts."

Moreover, it is interesting that one of the students mentioned that using perfect tenses rewarded more points, which is why she tried to include more perfect tenses in her essays.

Q3. What do you think about the feedback you got from Criterion? How was it different from the teacher feedback? Which one do you prefer – Criterion feedback, teacher feedback, or combined automated-teacher feedback?

As there were a few parts to this question, the answers were provided in separate tables.

Table 18. Students' Preferred Feedback Type

	Criterion Feedback	Teacher Feedback	Combined
	0	3 + 9	6 + 9
Total	0	3	15
Percentage	0	16.67%	83.33%

Table 18 indicates the number of students who preferred Criterion feedback, teacher feedback, and combined automated-teacher feedback. As illustrated in the table, no students preferred Criterion feedback alone but most of them (n=15, 83.33%) preferred combined automated-teacher feedback explaining the advantages and drawbacks of all feedback types in the following part of the question. Three students preferred teacher-only feedback, while six students preferred combined automated-teacher feedback directly. However, a group of students (n=9) thought they only had two options (either the Criterion feedback or the teacher feedback) and told before the researcher completed the question directly that they preferred teacher feedback. When they were told that combined automated – teacher feedback was also possible, they changed their minds and expressed their preference towards the combined automated – teacher feedback situation. In the end, while only three students (16.67%) chose to have feedback from the teacher, the rest of the students (n=15, 83.33%) preferred the combined automated – teacher feedback.

The reasons why the students preferred these types of feedback were provided in Table 19 from the responses they gave to the second half of the third question. As the students compared both types of feedback while explaining the reasons behind their preference, the themes emerged as "advantages and disadvantages of the Criterion feedback" and "advantages and disadvantages of the teacher feedback."

Table 19. Reflections on Different Types of Feedback

Themes	Codes	*f*	%
Advantages of the Teacher Feedback	Specific and In-depth feedback	8	10.52%
	Detailed feedback on the essay structure	4	5.26%
	Human touch	2	2.63%
	More focus on content	3	3.94%

	Focus on both strengths & weaknesses	1	1.31%
Total		18	23.68%
Disadvantages of the Teacher Feedback	Unnoticed mistakes	2	2.63%
	Oral feedback (not permanent)	1	1.31%
	The teacher's unavailability	2	2.63%
Total		5	6.57%
Advantages of Criterion Feedback	A clear indication of the mistakes	2	2.63%
	Detailed feedback	7	9.21%
	Focus on the form	5	6.57%
	Written feedback (permanent)	2	2.63%
	Feedback on mechanics	2	2.63%
	Color-coded feedback	2	2.63%
	Availability	1	1.31%
Total		21	27.63%
Disadvantages of Criterion Feedback	Superficial & General	9	11.84%
	Insufficiency	4	5.26%
	Difficult to understand	5	6.57%
	Highlights correct structures as mistakes	4	5.26%
	Insensitive to some mistakes	1	1.31%
	Misunderstands the content	2	2.63%
	Generic feedback on essay structure	4	5.26%
	No feedback on the content	3	3.94%
Total		32	42.10%
Total		76	100%

Although the majority of the students preferred combined automated-teacher feedback, they mentioned the disadvantages of Criterion feedback much more frequently (42.10%) than the disadvantages of teacher feedback (6.57%). Some of the disadvantages of Criterion mentioned by the students are as follows:

"Criterion is quite superficial in the mistakes it highlights. The feedback is too general and does not provide enough details to make us correct the mistakes properly. In other words, it does not provide necessary warnings to enable correction."

"The feedback provided by Criterion was difficult to understand. Sometimes I corrected a mistake highlighted by Criterion, but it still highlighted it as a mistake. At that point, I did not know what to do."

"I did not understand the color-coded feedback on organization because it was always there. It told me to write a thesis statement, but I already had it. It was like this for all the paragraphs, and I did not understand what the mistake was."

The advantages of both teacher and Criterion feedback were mentioned almost equally, 23.68% and 27.63%, respectively. The drawbacks of one feedback type were compensated for by the other, as shown in Table 19. The advantages expressed are as follows:

"Criterion may understand what I am trying to say in a wrong way. However, the teacher understands what is meant by the sentence and make comments accordingly."

"Criterion gave very good feedback on grammar and mechanics; it did not ignore even the smallest mistakes."

"Criterion feedback was very good on visible, simple mistakes while teacher feedback focused on organization, how to start a paragraph, what kind of paragraphs should be added, etc."

Because of these reasons, combined automated-teacher feedback was the best option, and most students preferred it. Criterion feedback was mentioned more than instructor feedback during the interviews since it was the new feedback type for the students, and they were excited to share their opinions about it.

Q4. To what extent do you think using Criterion improved your writing skills? In which area do you think you improved the most?

There were two parts to this question. To answer the first part, all the students who participated in the interviews (n=18) thought that Criterion improved their writing skills and performance.

Table 20. Reflections on Improved Writing Skills

Themes	**Codes**	***f***	***%***
Lower-level skills	Improved grammar	8	17.02%
	Improved punctuation	4	8.51%
	Improved vocabulary	7	14.89%
	Improved spelling	1	2.12%
	Improved sentence construction	4	8.51%
Total		24	51.06%
Higher-order skills	Improved text organization	7	14.89%
	Improved the quality of content	2	4.25%
	Improved use of conjunctions	2	4.25%
	Gained competence in essay writing	7	14.89%

Total		18	38.29%
Awareness of Mistakes	Being aware of basic mistakes	5	10.63%
Total		31	100%

Table 20 presents the areas the students said they improved. More than half of the answers (51.06%) included notions related to the improvement in lower-level skills. The students expressed that they improved, especially in grammar, vocabulary, and punctuation specifically. Four students especially mentioned how much improvement they observed in their use of punctuation because, in traditional writing classes, punctuation was not emphasized that much, and teachers may sometimes miss indicating the mistakes related to punctuation. The following statements include students' ideas related to the topic:

"It helped me in correcting grammatical errors the most. It directly indicated my mistakes and helped me to understand and correct them. I tried not to repeat them in the following essays. It also helped me to learn how to use punctuation correctly. When the use of punctuation is instructed, it is impossible to understand. However, when you see your punctuation mistake highlighted, you can easily learn."

It is interesting to note that the percentage of the notions that included improvement in higher-order skills is very close to lower-level skills (38.29%). The students mentioned improvement in their organization, content, and use of conjunctions and gained competence in essay writing. This is important because using AWE tools usually helps improve lower-level skills. Having comments from the students related to improvement in their higher-order skills shows another aspect of Criterion that is worth using in writing classes. Following statements include these comments:

"I realized an overuse of conjunctions at the beginning of sentences in my essays when Criterion highlighted them. Thanks to this feedback, I improved the paragraph format and the organization."

"I have never written an essay before. I have studied at state schools all my life, and I am not very good at English. Now, thanks to Criterion, I learnt how to write an essay."

The final theme that emerged from question 4 was "awareness towards mistakes," and around 10% of the answers included comments related to this theme. Using Criterion

made the students well aware of the mistakes they were making and likely to make. For this reason, they became more aware and careful about the mistakes in their writings.

Q5. What are the strengths and weaknesses of Criterion?

Since this question focused on the strengths and the weaknesses of Criterion, themes from the students' responses emerged accordingly. Table 21 presents the codes and themes that belong to question 5.

Table 21. The Strengths and Weaknesses of Criterion

Themes	**Codes**	***f***	***%***
Weakness - Interface	Difficult to use & navigate	9	12.16%
	Unappealing interface	3	4.05%
Total		12	16.21%
Weakness - Feedback	Highlighting the same mistakes despite the correction	2	2.7%
	Insufficient feedback	2	2.7%
	Difficult to understand some feedback	9	12.16%
	Does not highlight mistakes at higher levels	1	1.35%
Total		14	18.91%
Strengths - Time	Immediate feedback	6	8.10%
	Saves instructor time	1	1.35%
	Availability	1	1.35%
Total		8	10.81%
Strengths - Interface	Highlighted and categorized errors	7	9.45%
	Easy to use	9	12.16%
	Seeing both drafts on the same page	1	1.35%
	Shows grammar mistakes directly	1	1.35%
Total		18	24.32%
Strengths - Efficiency	Enabling learners to write many drafts	6	8.10%
	Providing adequate feedback	3	4.05%
	Leading to permanent learning	2	2.7%
	Providing hands-on learning experience	1	1.35%
	Improving writing	5	6.75%
	Being effective and efficient	5	6.75%
Total		22	29.72%
Total		74	100%

The table illustrates that the students mostly talked about the strengths of Criterion (64.86%) while they mentioned weaknesses less than that (35.13%). Criterion's weaknesses were mainly related to the interface of the website and the feedback it provided. Most students found the website difficult to navigate, especially at the beginning of the treatment; they had difficulty finding their essays, finding the results and the feedback, and copying and pasting a new draft despite the training done just before the intervention. However, four students who mentioned the difficulty related to the use of the website were talking about the difficulty of getting used to the website in the beginning. They admitted that it was pretty straightforward to use when they got used to the website. The weaknesses related to the feedback included Criterion's failure to identify mistakes correctly and the students' inability to understand some feedback because the students' English level was insufficient, and they did not know and were not used to the metalanguage. The weaknesses related to Criterion were expressed in the following statements:

"In the beginning, I could not find where to click to find my essay and write a new one. Later on, I figured out how to use the website."

"I did not understand the feedback in the beginning. I did not know the abbreviations provided as the feedback, I just saw them when the teacher trained us, but I forgot them. I learnt them better thanks to teacher feedback and Criterion feedback in the following essays."

As for the strengths of Criterion, the general themes that emerged were time, interface, and efficiency. It is interesting that interface was categorized both as a weakness and a strength. The interface of the website had its own advantages and disadvantages. While the disadvantages were mentioned above, the advantages that the students loved included the highlighted and categorized errors and seeing both drafts side by side on the same page. In addition, half of the students (n=9) expressed that the website was easy to use. Within the theme "strengths – time," the students expressed immediate feedback as one of the best sides of Criterion, as they liked having feedback while the ideas were fresh in their minds. Moreover, the theme of efficiency included many different codes as students expressed the efficiency of the program in various ways. Criterion was regarded as efficient because it allowed the students to write multiple drafts, hence practicing more. In addition, the students liked the feedback provided by Criterion because using Criterion, together with the feedback it provided, led students

to improve their writing and learn how to write an essay. Ultimately, it was a hands-on learning experience that resulted in permanent learning. The student expressions are as follows:

"The best thing about Criterion was the immediate feedback."

"I really liked the immediate feedback, and it does not need any time to process the essay; it directly provides the score and the feedback within seconds. I also liked to see the errors highlighted and categorized; this is how I learnt in which area I made the most mistakes."

"Criterion was available to me all the time, and I logged in to the website whenever I wanted to correct my mistakes. As a result of this, I had all my writings there to check, and it enabled me to learn better and improve my writing."

Q6. How did using Criterion affect your motivation?

All the students who participated in one-on-one interviews (n=18) asserted that using Criterion definitely motivated them in their writing class. The students elaborated on the question and explained the specific things related to Criterion that motivated them, and Table 22 presents the things that motivated the students.

Table 22. Motivational Factors Related to Criterion

Themes	**Codes**	***f***	***%***
Benefits of Automation	The chance to write multiple drafts	5	11.11%
	Immediate feedback	5	11.11%
	Autonomous feedback process	2	4.44%
Total		12	26.66%
Scoring	Getting high grades	4	8.88%
	Being rated by a program	4	8.88%
	The effort to increase the score	5	11.11%
Total		13	28.88%
Emotional Factors	Confidence	8	17.77%
	The chance the revise before the instructor	3	6.66%
	Being excited	3	6.66%
	Enjoyable process	4	8.88%
	Scared of writing before Criterion	2	4.44%
Total		20	44.44%
Total		45	100%

One of the best things students found motivating related to Criterion was the automation. The first one was the immediate feedback it provided within seconds. It is not possible for a human to detect all the grammar mistakes and provide feedback within seconds. The students liked to get feedback immediately while all the ideas were fresh in their minds. In addition, Criterion allowed them to write as many as ten drafts per essay, and they could practice as much as they wanted. In traditional writing classes, as teachers are constrained by time, they can usually give feedback on a maximum of two drafts. All errors must be corrected, and all improvements in all areas must be made within these two drafts. Also, teachers avoid giving essay assignments frequently because of the time limit. When Criterion is used, the only effort teacher must make is to create an assignment on Criterion. The students expressed that using Criterion as part of their writing class motivated them because of immediate feedback and the chance to write many drafts. The statements below include these ideas:

"Criterion gave us a chance to write ten drafts for one essay, so I had the chance to correct my mistakes, which motivated me."

"Because Criterion provides feedback immediately, you do not have to wait. You can correct your mistakes immediately while everything is fresh in your mind, which is motivating."

Another thing that motivated the students during the process was the scores provided by Criterion. The students liked to have higher scores from Criterion (n=4, 8.88%), they liked being rated by a program (n=4, 8.88%), and most importantly, the effort they showed to increase their score motivated them the most (n=5, 11.11%). They loved the idea of getting scores for each draft from a program, and this situation increased their motivation, and they mentioned those in the following statements:

"The teacher told us to increase our scores by at least one point. This motivated me, and I wanted to get high scores when I realized I could."

"I liked the idea of being evaluated by a program."

Finally, the students mentioned how they felt during the process. In the beginning, some students were excited about using a new program (n=3, 6.66%), and others were scared of writing (n=2, 4.44%) in general. However, in the end, after the treatment, almost half of the students (n=8, 17.77%) felt more confident in their writing performance; some of the students (n=4, 8.88%) enjoyed it a lot. In addition, some

students (n=3, 6.66%) expressed that they felt more comfortable because they had the chance to correct some of their mistakes before the instructor saw their paper. Their ideas were reflected in the following statements:

"Now, I think I am better at writing; as a result, I feel more comfortable in the exams that include writing. Also, correcting the mistakes, especially grammar, was quite enjoyable."

"Before we started writing the essay, I was anxious and scared about writing. In our writing classes this semester, I had a chance to write many essays with many drafts. Therefore, I am now less anxious and less scared, and I have become a better writer."

"Criterion made writing an enjoyable process; it basically gamified everything."

Q7. Do you have any suggestions related to the whole process or the use of Criterion?

Most of the students (n=11) expressed that they were content with the way everything was conducted; they liked the fact that they used Criterion as part of their writing classes, and they also liked to get combined automated-teacher feedback. Some students (n=6) asserted that they were happy to use Criterion during this semester, some students (n=2) were happy about the number of essays they wrote, and some other students (n=3) were happy that they had the chance to write more thanks to Criterion. A few students are so content with Criterion that they want to use it when they start their departments and suggest it to their friends. The students' contentedness can be observed from the following statements:

"I am quite happy with everything. We wrote an essay every week during the intervention. I am so happy because we did not have this opportunity in the fall semester."

"I am so happy we used Criterion in our writing class this semester. When I talked to my friends from other classes, I understood that we did a lot of writing practice compared to them. We got used to essay writing, and now I can easily write an essay."

However, there were a few things that the students did not like about Criterion, which were related to the topics. The first area of discontent was related to the fact that Criterion gave lower scores when the topics were assigned by the teacher, not chosen

from the topic library. Two students mentioned that. Another two students found the topics chosen from the topic library difficult, and they mentioned that the topics required much research before they started writing their essays. In addition to these comments, students expressed their suggestions to answer question 7, and the responses are given in Table 23. Almost every student who said they had a suggestion expressed a different idea, so two themes emerged from the responses, and they are given in Table 23.

Table 23. Suggestions Related to the Whole Process or the Use of Criterion

Themes	Codes	*f*	*%*
Criterion-related suggestions	The chance to use Criterion individually	2	16.66%
	Direct feedback on grammar mistakes	1	8.33%
	Being free in choosing essay topics	1	8.33%
	More interactive platform	1	8.33%
Total		5	41.66%
Process-related suggestions	Getting teacher feedback earlier	1	8.33%
	Having group practice at the beginning	1	8.33%
	Using Criterion under exam conditions	2	16.66%
	Having more practice	1	8.33%
	Having more time to write essays	1	8.33%
	Using thesaurus tool	1	8.33%
Total		7	58.33%
Total		12	100%

The suggestions embodied both the whole feedback process and the use of Criterion as part of the writing class. Some of the suggestions (the chance to use Criterion individually, being free in choosing essay topics, direct feedback on grammar mistakes, and a more interactive platform) were related to the limitations of the website. They can be offered as a suggestion from the user experience to the developers of Criterion. Some other suggestions (having more time to write essays and using the thesaurus tool) could have been made but were not preferred by the instructor. The students could have been given more time to finish their essays, but they would not get used to writing timed essays, which they would do at the end of the semester in the English Proficiency Exam (EPE). The thesaurus tool was specifically turned off by the instructor because the students' level of English was not enough to notice nuances between the vocabulary suggested by the Criterion thesaurus tool.

Other suggestions, such as getting teacher feedback earlier and having a group practice in the beginning, could have been implemented. A pilot study could have been conducted to explore the Criterion and help students get used to the program if we were not limited by the curriculum. Teacher feedback was not given earlier as the researcher wanted the students to take as much time as they wanted to write multiple drafts. The codes and themes that emerged from the student responses in one-on-one interviews were provided in this section. The analysis and discussion related to these themes will be presented in the next chapter.

5. CONCLUSION

This chapter consists of five parts: discussion of the findings regarding each research question, conclusion, pedagogical implications, limitations, and recommendations for future research. In the first part, the results obtained in the present study were examined individually in relation to each research question. Then, following the conclusion part, pedagogical implications for EFL classrooms were provided. In the two final sections, the limitations of this study were mentioned, and recommendations for further studies were presented.

5.1. Discussion of the Findings

This section provides a brief overview of the findings from the data analysis, and the outcomes were examined in relation to each research question.

5.1.1. Discussion on the Findings for RQ1

5.1.1.1. Difference between groups

The first research question was, "Do the use of Criterion and receiving combined automated-teacher feedback affect Turkish EFL students' writing performance significantly compared to teacher feedback only?" To be able to answer this question, several statistical tests were conducted, and the results of the descriptive and inferential statistics were presented in the results chapter. In this section, a discussion of these results pertaining to descriptive and inferential statistics will be provided.

After making a comparison between the post-test scores of the experimental and control groups, it was discovered that there was no significant difference in the writing proficiency of the two groups following the intervention ($U=186.500$, $p=.759>.05$). In addition, the gain scores for both groups were calculated. The results of the Mann-Whitney U Test yielded no significant difference between the two groups in their gain scores. This suggests that the combined automated-teacher and full-teacher feedback

used in the study had a similar effect on the students' writing performance, as demonstrated by their post-test scores.

The results align with the literature; as Wilson & Czik's study (2016) illustrates, the students who got feedback from their teachers and an AWE tool improved their writing persistence. However, there was no significant difference in the final-draft writing quality. A similar study was conducted in a Turkish context with tertiary-level students, with experimental and control groups. In that study, while the experimental group was in an automated-teacher feedback condition, the control group obtained teacher feedback only. The results indicated no significant difference in the writing performance of the two groups in their final writings (Han & Sari, 2022). Another study conducted with university students with lower-intermediate English levels concluded that integrating an AWE tool into the writing classes did not significantly improve students' writing (Huang & Renandya, 2020). However, it is essential to note that Huang and Renandya's (2020) study employed AWE and peer feedback; no teacher feedback was provided.

There are not many studies that yield a similar result. It can stem from the fact that the present study employed a combined automated – teacher feedback condition in the experimental group and teacher condition in the control group. However, many studies in the literature either did not have control groups (e.g., Chen & Cheng, 2008; Ebyary & Windeatt, 2010; Jinla, 2014; Li et al., 2015; Parra & Calero, 2019; Scharber et al., 2008; Wali & Huijser, 2018) or control groups did not get any feedback (e.g., Cheng, 2017; Kellogg, Whiteford & Quinlan, 2010). In addition, among the studies chosen for Stevenson & Phakiti's (2014) review, most of them had a control group with no feedback condition (e.g., Franzke, Kintsch, Caccamise, Johnson, & Dooley, 2005; Grimes, 2008). They assert that for the control groups in those studies, one of the following three options happened: the students in the control groups did not write anything, they wrote essays, but they did not get any feedback, or they got feedback as in the traditional writing classes, but it was not mentioned in the study. Moreover, these studies reported the positive effects of AWE tools on the students' writing performance. However, it is unknown whether these effects were significant because of the inequality of the conditions for both groups. Furthermore, the studies that compared the performance of both groups under the conditions mentioned above (e.g., Rock, 2007; Schroeder, Grohe, & Pogue, 2008) avoided reporting the facts and

sugarcoated the results of the study, indicating that the students in the experimental group improved significantly in their writing performance.

When this is the case with many studies in the literature, the feedback conditions employed for both groups were too strong in this study, and, understandably, there is no significant difference between the two groups. While the experimental group received automated feedback a few times for a single essay and teacher feedback on the final draft, the control group got teacher feedback on both drafts. Moreover, when Tables 8 and 12 are checked, it is observed that the mean scores of the pre–test (Exp: 3.46; Cont: 3.93) and the post–test (Exp: 4.25; Cont: 4:43) were higher for the control group. This might mean that better students were clustered in the control group, and that is why there was not a significant difference between the two groups despite the amount of writing (Exp:~350 drafts; Cont: 160 drafts) the experimental group did.

On the other hand, there are two studies with similar research designs. In one study, Choi (2010) found out that the students who used Criterion together with teacher feedback outperformed those who did not use Criterion or used Criterion optionally regarding holistic quality and accuracy in grammar. In addition, Tang and Rich (2017) reported similar results: the learners in the combined automated-teacher feedback condition improved their writing skills better than those in the teacher feedback condition. Although Choi (2010) explained that teacher feedback was available to students more than once, in every step, from planning to the final draft, the feedback details were not provided in both studies. It is also significant to indicate that the groups in Choi's study (2010) were from two different institutions: one in the US and one in South Korea, and different instructors taught and gave feedback to the students. Although Choi (2010) asserted that the conditions were identical for all groups, the difference may stem from the difference in the conditions of the two groups.

As a result, the results of the present study emphasize the significance of research that utilizes stronger counterfactuals that are more aligned with real-world conditions, which is providing teacher feedback to the control group. Such research is crucial to gain a more accurate understanding of the actual impact of automated feedback. This impact is likely to be lesser than that reported in previous research that utilized no feedback and no control group conditions (Wilson & Czik, 2016).

5.1.1.2. Difference within groups

Descriptive and inferential statistics were conducted to see whether the experimental and the control groups improved their writing performance throughout the intervention. After making a comparison between the pre- and post-test results of the control group, the results showed no significant difference between the scores of pre- and post-tests ($z=-1.315$, $p=.189>.05$), which means the students in the control group did not improve much before and after the intervention. Choi's (2010) research illustrated similar results. There was no notable improvement in the students' writing proficiency in the teacher feedback condition due to the fact that teachers were unable to provide timely and adequate feedback on the learners' essays due to the large class size and time constraints. There are not many studies that support the findings of the present study for the control group because many studies in the literature either do not have a control group (e.g., Chen & Cheng, 2008; Ebyary & Windeat, 2010; Jinla, 2014; Li et al., 2015; Parra & Calero, 2019; Scharber et al., 2008; Wali & Huijser, 2018) or control groups did not have any feedback at all (e.g., Cheng, 2017; Kellogg et al., 2010). When this is the situation, there were no pre-tests and post-tests or interventions for the control group, and no results related to control groups were yielded.

On the other hand, after making a comparison between the pre- and post-test results of the experimental group, the results showed a significant difference between the scores of pre- and post-tests ($z=-2.642$, $p=.008<.05$), which means the students in the experimental group improved in their writing performance significantly at the end of the intervention. There are many studies that are in line with this result; they concluded that the students who used an automated writing evaluation tool in their writing classes or the learners who obtained combined automated–teacher feedback improved their writing performance. The results from Ebyary and Windeatt's (2010) study indicated a steady increase in the holistic scores of students from their first essay to the last. According to Cheng (2017), there was a significant increase in scores observed across reflective journals of the experimental group. Furthermore, the students in the combined feedback condition showed the most improvement, according to Mohsen & Alshahrani's study (2019). In Wilson and Czik's (2016) study, the students who got feedback from their teachers and an AWE tool improved their writing persistence. In Tang and Rich's study (2017), the students in the experimental group improved much more than those in the control group from the pre-test to the post-test. In addition, the

students who employed an AWE tool in their writing classes enhanced their writing scores by one point on a six-point scale across subsequent drafts (Foltz, Lochbaum, & Rosenstein, 2011 as cited in Foltz, Streeter, Lochbaum, & Landauer, 2013) and got higher scores from both the AWE tool and the human rater than those who did not practice with Criterion. However, these students received no teacher feedback (Wang et al., 2013).

Many studies indicate improvement in different areas of writing, such as writing complexity and fluency (Link et al., 2020), writing accuracy (Dikli & Bleyle, 2014; Hartshorn et al., 2010; Li et al., 2015; Lv, 2018), and lower-level writing skills, such as grammar, vocabulary, spelling, and punctuation (Barrot, 2021; Ghufron & Rosyida, 2018; Jayavalan & Razali, 2018; Parra & Calero, 2019; Wali & Huijser, 2018) for the students who use automated writing evaluation tools as part of their writing instruction.

The results for the experimental group contradict the results of Choi's (2010) study, in which he found out that the integration of an automated tool in the writing classes did not affect the improvement of students' writing performance.

5.1.2. Discussion on the Findings for RQ2

The second research question was "What are the experimental group students' perceptions of combined-automated teacher feedback and their attitudes towards using Criterion as part of their writing classes?" To be able to answer this question, one-on-one semi-structured interviews were conducted with 18 students from the experimental group. Each student was asked seven questions (Appendix 2). The responses to each question were analyzed and appropriately coded, and themes were created.

To begin with, the students expressed their trust in Criterion scores. However, some asserted that the scores are only trustworthy when the topics are chosen from the Criterion Topic Library. Otherwise, Criterion gives lower scores to the essays whose topics are provided by the instructor because automated writing evaluation tools must undergo training using a vast collection of pre-scored essay samples to evaluate student essays accurately (Burstein, 2003; Elliot, 2003; Landauer et al., 2003). For Criterion, the topics in the Criterion Topic library are pre-scored prompts, and it is not equipped to score essays with unfamiliar prompts as it lacks the necessary training to evaluate them (Dikli, 2006). This situation is a drawback for the automated writing evaluation tools as detected by the students. During the intervention, the researcher employed the

topics from the Criterion topic library to avoid this situation. To be able to have the correct essay type – persuasive – on Criterion, TOEFL topics were assigned to the students from the library. Then, the students mentioned during the interviews that the topics were too difficult and too abstract to write about because their level of English is lower (A2); they needed more concrete topics to write their essays about. However, previous research shows that some of the students who used an AWE tool as part of their writing classes did not trust the scores provided by the AWE tools, as there were discrepancies between the teacher score and the automated score (Chen & Cheng, 2008; Li et al., 2014), some students trusted it and some instructors still wanted to benefit from it (Grimes & Warschauer, 2010; Li et al., 2014).

Related to the reliability of the holistic scores of Criterion, there are a few things that the students mentioned. It is interesting to note that while a few students mentioned that Criterion rewards longer essays and they add sentences to increase their score, one student mentioned that he just wrote 180 words (should have been at least 250) and still got a good grade. The literature also supports that e-rater is inclined to value long essays more than human raters (Bridgeman, 2013; Chen & Cheng, 2008; Wang et al., 2013). Moreover, teacher observations in the literature support these results; they expressed that AWE favors length (Li et al., 2015). Favoring certain expressions, being insensitive to content and some mistakes, and the impact of basic mistakes on the final score are the other issues reported by the students related to the inaccuracy of Criterion scores. The research by Powers et al. (2001) supports these comments by adding that one of the most significant concerns related to AEE is the possibility of inappropriately rewarding or punishing particular aspects of examinees' writing, which poses a danger to its validity.

During the intervention, the students in the experimental group were encouraged to increase their scores given by Criterion at least one point across subsequent drafts. The results of each draft indicated that every student managed to increase their score. The students were asked how they managed to increase their scores and what they did exactly.

76.36% of the responses indicated that the students corrected the mistakes highlighted by Criterion and increased their score as a result. These highlighted mistakes were on lower-level features such as grammar, vocabulary, punctuation, spelling, and organization, as the argument goes that AES focuses on the text's surface grammatical

characteristics and semantically important vocabulary and rates the essays accordingly (Huang, 2014). Focusing on the mistakes highlighted by Criterion and correcting them worked for the majority of the students, and they increased their scores by at least one point. Teacher observations from the literature support this; according to them, correcting minor grammatical mistakes can improve the total score (Li et al., 2015).

There were other strategies mentioned by the students to increase their Criterion score and hence improve their writing. The strategies included increasing word count, rewriting some parts, studying sentence structures and vocabulary, doing research on the essay topic, using a variety of tenses, studying with a friend, improving reading skills, and spending time on the later drafts. It is valuable that the students employed and mentioned these extra strategies to increase their Criterion score and improve their writing because the findings from Ebyary and Windeatt's (2010) study contradict that. They concluded that although AWE programs may assist students in rectifying grammar and mechanics errors (e.g., spelling and capitalization), they typically do not make any revisions like changes to collocations or sentence structure, even though they submit many drafts.

When they were asked about their feedback preference, most of the students preferred combined automated-teacher feedback, whereas none of the students preferred automated feedback situation. The previous research is in line with these results. Many studies (Chen & Cheng, 2008; Choi, 2010; Link et al., 2020) preferred to use AWE tools together with the teacher feedback, not alone. The results from Chen & Cheng's (2008) study demonstrate that the AWE feedback was unable to address students' writing issues sufficiently and in a meaningful way and cannot get what the students truly meant as teachers did (Wang et al., 2013). For this reason, teacher feedback is also required (Attali et al., 2013; Benali, 2021; Li et al., 2014; Wali & Huijser, 2018). Some students asserted that having AWE feedback without complementary teacher feedback negatively affected their attitude towards the AWE tool and the writing class, and the students in the combined feedback condition showed the most improvement (Mohsen & Alshahrani, 2019). These findings from the literature were the reasons why combined automated–teacher feedback was employed in this study, and the students liked it.

Only three students preferred teacher feedback noting that they like the "human touch" and the focus on the strengths and weaknesses. These comments align with the position

statements on writing assessment by Conference on College Composition and Communication (2004, 2009) as they asserted that humans are social creatures and writing is a social construct, and using AES systems to evaluate writing dehumanizes the writing construct. In the study by Huang (2014), the students emphasized the interaction they get from the teacher and the importance of negotiation, which is an essential part of the writing process. In addition, AWE systems lack the depth of any genuine interaction that may occur in the actual world, and the students regard the AWE tools negatively in that score (Chen & Cheng, 2008; Wang et al., 2013).

In addition, the advantages of teacher feedback mentioned by the students included specific, in-depth, and detailed feedback on the essay structure with a focus on content. The ideas in these responses are supported by the previous research. Human feedback is characterized by its specificity, attention to detail, and ability to discern what actions are necessary for the students to enhance their writing (Chen & Cheng, 2008; Huang, 2014). Instructor feedback can encourage the improvement of the writing construct in the areas such as writing complexity and fluency (Link et al., 2020) by providing detailed feedback on content, development, coherence, and cohesion (Huang, 2014). As a result, the logical structure of arguments can best be evaluated by humans (Bridgeman et al., 2012; Gao, 2021). Finally, the research by Zhang and Hyland (2018) demonstrated that skilled ESL teachers could provide more comprehensive corrective feedback on student writing.

The students offered reasons why they preferred a certain type of feedback, usually by comparing two types: automated and teacher feedback. The teacher's superiority in giving feedback on content was compared to Criterion's advantage of giving detailed feedback on form and mechanics by clearly indicating the lower-level mistakes. Similar results were presented in the literature. Whereas AES focuses on the text's surface grammatical characteristics and semantically important vocabulary and rates the essays accordingly, human raters typically look at other aspects of writing besides grammar and vocabulary, such as content, development, coherence, and cohesion (Huang, 2014). Moreover, Chinese EFL learners discovered that AWE feedback could identify spelling mistakes and the majority of lexical mistakes correctly (Jingxin & Razali, 2020). The instructors in Li et al.'s (2015) study also reported their belief in the positive effect of AWE corrective feedback in improving grammar and mechanics, and the students displayed a positive attitude in terms of the AWE tool helping them

improve their writing accuracy (Dikli & Bleyle, 2014; Lv, 2018) thanks to the reasons the students in this study expressed.

The students reported that while some mistakes went unnoticed in the teacher feedback, Criterion was consistent and persistent in the feedback it gave. These comments are supported by the literature. In Almusharraf & Alotaibi's (2022) study, Grammarly detected more errors than human raters. However, all these identified errors were related to style, word choice, format, and punctuation and were repeated throughout the texts. However, there were other findings that contradicted these comments. For example, Tömen (2022) illustrated that the feedback provided by Criterion on grammar and mechanics was similar to teacher feedback, and the number of errors identified by the teacher and Criterion was very close. Moreover, human feedback on lower-level features was still more than the AWE tool with better quality-more accurate and focused (Chen et al., 2009; Dikli, 2010; Dikli & Bleyle, 2014; Link et al., 2020) and as in Dikli and Bleyle (2014) and Park's (2019) studies, Criterion and Grammarly, respectively, identified fewer errors than human raters. However, AWE is trusted with its feedback on lower-level features, and the instructor in this study preferred to focus on higher-order features and sometimes ignored the lower-level features thinking that Criterion deals with that part. This is the reason why students mentioned this.

The disadvantages related to Criterion mentioned by the students included the feedback being superficial, general, insufficient, difficult to understand, and highlighting correct structures as mistakes, such as highlighting the 'title' of an essay as a fragment error. The students in Han and Sari's (2022) study also mentioned the same problem. Criterion feedback sometimes fails to identify some mistakes and gives generic feedback on essay structure. As for the content, it either misunderstands the content, or there is no feedback related to it. Previous research aligns with these answers. As mentioned in the literature, the feedback from an AWE tool can be unfocused, indirect, vague, formulaic, repetitive (Chen & Cheng, 2008; Wang et al., 2013), not specific, long (Dikli, 2010), non-dialogic and fallible (Bai & Hu, 2017) and it also ignores coherence and richness of the content (Chen & Cheng, 2008; Wang et al., 2013). For example, Criterion may suggest revising a punctuation error by highlighting the whole sentence (Link et al., 2020), or it indicates that the sentence is

wrong but does not provide further details related to the error, and this situation frustrates the students (Li et al., 2015).

The final comparison made by the students is in terms of availability. While Criterion is available 24/7, this is not possible for a teacher. Providing feedback through AWE offered clear benefits compared to feedback given by teachers, particularly in terms of speed, ease of access, and the ability to revise multiple times (Chen & Cheng, 2008; Dikli, 2006).

All the students who participated in the interviews expressed that using Criterion in their writing classes improved their writing skills. Improvement in lower-level skills such as grammar, punctuation, vocabulary, spelling, and sentence construction was mostly mentioned by the students. The participants specifically mentioned how well Criterion's feedback on punctuation was and how much they improved in terms of their use of punctuation. The literature is also in line with these responses. Many previous studies on the use of AWE tools as part of writing instruction discovered that AWE tools improved students' lower-level writing skills, such as grammar, vocabulary, spelling, and punctuation (e. g. Barrot, 2021; Ghufron & Rosyida, 2018; Jayavalan & Razali, 2018; Parra & Calero, 2019). In their studies, Wang et al. (2013) and Liao (2016) investigated the effects of an AWE tool on writing accuracy. The results of the post-tests showed that the students in the experimental group were significantly more accurate in capitalization, punctuation, the use of articles, tenses, different vocabulary, etc., than the control group. In addition, the students displayed a positive attitude regarding the AWE tool, helping them improve their writing accuracy (Dikli & Bleyle, 2014; Lv, 2018). However, there are some studies reporting that AWE tools did not improve students' holistic writing quality (Huang & Renandya, 2020; Wilson & Roscoe, 2020) because of many issues in grammar feedback.

The literature insists that Criterion is not efficient in giving feedback on higher-order skills. In Chukharev-Hudilainen and Saricaoglu's (2016) study, both the students and the instructors were not satisfied with the feedback on content and organization, and Hoang and Kunnan (2016) reported that the AWE tool was insufficient in identifying organizational issues at the discourse level and off-topic content. However, the students in this study were satisfied with the feedback they received from Criterion. They reported that they improved in higher-order skills such as text organization, the quality of content, and the use of conjunctions and gained competence in essay writing

in general. It was emphasized in literature as a negative aspect of AWE feedback that there was too much emphasis on the use of transition words (Chen & Cheng, 2008; Wang et al., 2013), yet, the participants in this study regarded it as something positive.

There is another aspect to this question, which is awareness. The students reported that using criterion made them more aware of their own mistakes and made them more careful in their subsequent writings. These comments were also supported in the literature. Grammarly provided the students with the ability to notice, which plays a vital role in writing development. As a result of the metalinguistic explanation provided by Grammarly, students could notice the errors they made (Barrot, 2021).

The weaknesses related to the use of Criterion in the writing class emerged under two themes: interface and feedback. While many studies mention weaknesses related to the feedback from AWE tools, not many studies refer to the weaknesses or strengths related to the interface to the best of the researcher's knowledge. In this study, participants mentioned having difficulty navigating the website, especially in the beginning. However, Criterion was very easy to use once the participants got used to it. Whereas the students hated the unappealing interface, they loved to see errors highlighted and categorized above, together with the ability to see both drafts on the same page. It is important to note that in Han & Sari's (2022) study, the students reported encountering minor difficulties during their initial use of the program. However, they managed to resolve these issues quickly.

The weaknesses related to the feedback provided by Criterion include highlighting the same mistakes despite the correction, insufficient feedback, and difficulty understanding some feedback. The students had difficulty interpreting the feedback, especially in the beginning, and the feedback was in English, and the metalanguage was a problem for some students. Moreover, the generic, color-coded feedback on the essay structure was meaningless for some students; they did not understand what Criterion meant by that feedback. Previous studies align with these results. The primary issues related to the grammar feedback from the AWE tools were identified as the limited range of error types detected, and the inconsistency in the way errors were flagged (Chen & Cheng, 2006). The same situation occurred with Criterion feedback in which some of the identified error codes were correct and some were not and most of the grammatical errors went unidentified (Han et al., 2006; Lavolette et al., 2015; Tetreault & Chodorow, 2008). In Li et al.'s study (2015), some of the

instructors raised concerns related to the quality of the feedback. Moreover, the feedback from the AWE tools is provided in English mostly, and this situation was frustrating for the EFL students at lower levels as they had difficulty understanding the metalanguage and the feedback well (Wang et al., 2013).

Criterion's strengths can be categorized under three themes: Time, interface, and efficiency. Strengths related to the interface were mentioned above. As for time, the students loved getting immediate feedback and having access to Criterion all the time. Immediate feedback was one of the best features of Criterion, mentioned by many students that participated in previous studies. Despite some drawbacks, they preferred AWE tools to get immediate feedback (Grami, 2020). It is asserted that providing immediate holistic and analytical feedback is an advantage of AWE tools (Dikli, 2006; Yeh et al., 2007), and many students expressed their satisfaction related to the AWE tool because of the immediate feedback. In addition, the students expressed that as Criterion dealt with feedback on lower-level skills, the teacher had more time to focus on the content and the logic in the essays written by the students. For example, Link et al.'s (2020) study found that the teacher gave almost half as much feedback when the AWE tool was part of the writing instruction. In line with Link et al.'s (2020) study, teachers in Wilson and Czik's (2016) study reported that using AWE tools saved them some time while giving feedback, as the feedback on lower-level writing skills was handled by the AWE tool.

Criterion was reported to be efficient as it enables learners to write many drafts; hence, they are encouraged to write more, practice more, and improve their writing. In addition, it leads to permanent learning by providing hands-on learning experience and adequate feedback. The second-best feature of Criterion the students expressed, in addition to immediate feedback, was the chance to write many drafts. These two features were also reposted by many students in previous research as the advantages of Criterion. AWE provides students with feedback for every draft they write in no time, along with feedback on errors and relevant explanations (Kern & Warschauer, 2000; Milton, 2006; Ware & Warschauer, 2006). Providing immediate and constant corrective feedback as part of writing instruction encouraged students to write more and improve their writing accuracy (Hartshorn et al., 2010; Li et al., 2015). Chen and Cheng (2008) and Dikli (2006) concluded that the advantages of Criterion were speed, availability, the chance to write multiple drafts, and increasing learner autonomy.

The students who participated in the interviews all agreed that the use of Criterion motivated them. The literature also concluded that AWE systems led to more motivation in the writing class (Grimes & Warschauer, 2010; Warschauer & Grimes, 2008). They encouraged students to be more motivated to write and produce more and longer content (Li et al., 2014; Li et al., 2015; Wang, 2013). Although this situation can be attributed to the novelty effect (Chen & Cheng, 2008), it motivates students to some extent.

The factors that contributed to their motivation were recited by the participants. These factors were grouped under three themes: Benefits of automation, scoring, and emotional factors. The benefits of automation included the chance to write multiple drafts and immediate feedback, whose connection to literature was already mentioned. The students included these a few times in their responses as they liked these features a lot. The same theme included another code, which was an autonomous feedback process. The autonomous feedback process was closely related to the immediate feedback and the chance to write multiple drafts provided by Criterion. It made the students autonomous, freeing them from their dependence on the teacher. The idea that Criterion provides learner autonomy was supported by the literature. Using an AWE tool promoted learner autonomy with the resources it provided and the functions it offered (Dikli, 2006; Ghufron & Rosyida, 2018; Warschauer & Grimes, 2008). Moreover, elevated learner autonomy helped students to become more ready to produce more and better writing (Wang et al., 2013).

Another source of motivation was scoring, in which the students expressed how getting high grades, being rated by a program, and the effort to increase the score motivated them. In Han and Sari's (2022) study, most students believed that getting a score from the program was beneficial because it indicated their writing skills and inspired them to improve their writing to earn a better score. The students in this study specifically reported that they felt motivated the most because they were told to increase their overall score by the instructor by at least 1 point. According to a study by Li et al. (2014), establishing an overall benchmark score that students must attain before handing in their essays to their instructors could enhance learner interaction with AWE feedback. In addition, the participants in Zhang's study (2017) were incentivized by the possibility of achieving higher overall scores through multiple revisions. However, contrary to this study, they experienced a sense of discouragement when their scores

did not improve despite repeated revisions. In this study, the students did not report any discouragement because most of them said they managed to increase their scores by correcting the mistakes highlighted by Criterion.

The final source of motivation was how the students felt about the use of Criterion in their writing classes. The responses included negative feelings related to the writing process, which were being scared of writing and nervous about showing their mistakes to the instructor. These results can also be found in Han and Sari's (2022) study, in which the students conveyed that they felt at ease when writing on the program because it allowed them to rectify minor mistakes before submitting the essay for teacher evaluation. This is similar to Grimes and Warschauer's (2010) research, where the participants found machine evaluation less intimidating than human evaluation. Positive feelings as the source of increased motivation included increased confidence in writing and English in general, excitement to use the program, and enjoyment of the whole process. The findings from the literature are in line with these responses. A study by Wilson and Roscoe (2020) discovered that the use of an AWE tool as part of writing instruction improved the students' self-efficacy, and their scores on state tests have hence improved indirectly. Huang & Renandya (2020) reported that the students liked the feedback they got from Criterion and had a positive attitude towards the use of AWE tools in writing classes (Dikli & Bleyle, 2014; Lv, 2018)

Finally, the students expressed their satisfaction related to the use of Criterion and the whole writing process with combined automated – teacher feedback. In line with the student feedback, the literature suggests the combined automated – teacher feedback as the best option. For example, one teacher in Chen and Cheng's (2008) study depended on AWE feedback alone, but the students expressed their discontent at the end of the semester, mentioning that they needed teacher feedback to guide them. In addition, some students asserted that having AWE feedback without complementary teacher feedback negatively affected their attitude towards the AWE tool and the writing class (Mohsen & Alshahrani, 2019). It is interesting to note that contrary to what is suggested, Wali and Huijser (2018) used feedback from an AWE tool only, and more than half of the students reported that the AWE tool improved their grammar, content, and vocabulary. However, they still needed teacher feedback.

5.2. Conclusion

This mixed-method study was conducted at a state university with Turkish EFL learners in the writing classes of a preparatory program. The purpose was to see whether the students in the experimental condition in which the students used an automated writing evaluation program – Criterion – and obtained combined automated – teacher feedback to their writings would perform in their writings better than the students in the control condition in which they received their regular writing training and teacher feedback. The data to receive this information was collected through pre- and post-tests conducted in both groups. In addition, the attitudes and perceptions of the students who used Criterion and received combined automated–teacher feedback were explored via one-on-one interviews.

First, a comparison was conducted through statistical analysis to determine if one feedback model was superior to the other. However, the results showed no significant difference between the two feedback procedures in improving the learners' holistic writing scores. There was no significant difference between the two groups at the end of the intervention. Despite the small difference between the groups, the difference within the groups was analyzed. When the pre- and post-tests of the experimental and the control groups were checked, the significant difference within the group was observed only for the experimental group. It means that the students in the experimental group improved their writing from the pre-test to the post-test significantly. The descriptive statistics also show improvement in the writing skills of the control group; however, this improvement is not as significant as the inferential statistics show for the experimental group.

Second, the attitudes and perceptions of the students in the experimental group were investigated through seven questions (Appendix 2) in the interview. Several themes emerged from the student responses, most of which aligned with the literature. The students had a positive attitude towards the use of Criterion in class and received combined automated–teacher feedback. Most students preferred combined automated – teacher feedback (Attali et al., 2013; Benali, 2021; Chen & Cheng, 2008; Choi, 2010; Li et al., 2014; Link et al., 2020; Wali & Huijser, 2018). While teacher feedback is specific, in-depth, and detailed on the essay structure with a focus on content (Bridgeman et al., 2012; Chen & Cheng, 2008; Gao, 2021; Huang, 2014; Link et al.,

2020), Criterion feedback is detailed on form and mechanics and indicates the lower-level mistakes (Dikli & Bleyle, 2014; Huang, 2014; Jingxin & Razali, 2020; Li et al., 2015; Lv, 2018). However, the feedback from Criterion was criticized as being superficial, general, insufficient, difficult to understand, and sometimes fallible (Bai & Hu, 2017; Chen & Cheng, 2008; Dikli, 2010; Han & Sari, 2022; Li et al., 2015; Link et al., 2020; Wang et al., 2013) by the participants in this study. All the participants expressed that Criterion helped them to improve their lower-level writing skills (Barrot, 2021; Dikli & Bleyle, 2014; Ghufron & Rosyida, 2018; Jayavalan & Razali, 2018; Liao, 2016; Lv, 2018; Parra & Calero, 2019; Wang et al., 2013) and created awareness towards their own mistakes (Barrot, 2021).

The students mentioned the strengths and weaknesses that they observed related to Criterion. While the weaknesses were the interface (Han & Sari, 2022) and the feedback (Chen & Cheng, 2006; Han et al., 2006; Lavolette et al., 2015; Li et al., 2015; Tetreault & Chodorow, 2008; Wang et al., 2013), the strengths included time, interface, and efficiency. The strengths that were repeatedly mentioned by the students included immediate feedback (Dikli, 2006; Grami, 2020; Yeh et al., 2007) and the chance to write many drafts (Kern & Warschauer, 2000; Milton, 2006; Ware & Warschauer, 2006).

All the participants expressed that using Criterion motivated them (Grimes & Warschauer, 2010; Warschauer & Grimes, 2008). The reasons for this motivation were explained as the benefits of automation, scoring, and emotional factors. While the benefits of automation included the chance to write multiple drafts and immediate feedback, scoring included getting high grades, being rated by a program, and the effort to increase the score (Han & Sari, 2022; Li et al., 2014). Moreover, the students had a positive attitude towards the use of Criterion in general (Dikli & Bleyle, 2014; Huang & Renandya, 2020; Lv, 2018), and they were satisfied with receiving combined automated – teacher feedback (Chen & Cheng, 2008; Mohsen & Alshahrani, 2019).

Some responses that emerged in this study did not align with the previous studies. One of the responses that were not echoed in the literature was related to the scoring of Criterion, which rated accurately when the topic was chosen from the topic library but gave lower scores when a random topic was assigned.

One code that emerged from the responses of the participants in this study contradicted what was presented in the literature. The students in this study mentioned their extra efforts to improve their writing. In contrast, previous studies concluded that students made no revisions other than correcting the Criterion mistakes (Ebyary & Windeatt, 2010). Another code that surfaced in this study that contradicted what was presented in the literature was related to the amount of lower-level feedback the students received from the teacher. While some studies (Chen et al., 2009; Dikli, 2010; Dikli & Bleyle, 2014; Link et al., 2020; Park, 2019) concluded that human feedback is still more than AWE feedback on lower-level skills, the participants in this study disagree with that. In addition, the students reported that they improved in higher-order skills such as text organization, the quality of content, and the use of conjunctions as well as gaining competence in essay writing in general, contrary to what was presented in the literature (Chukharev-Hudilainen & Saricaoglu, 2016; Hoang & Kunnan; 2016).

5.3. Pedagogical Implications

Based on the results of this study, there are a few classroom implications that could be recommended to EFL writing instructors and school administrators. These implications can help all the stakeholders make informed decisions regarding the integration of technology in writing instruction and the implementation of better writing assessment techniques in their educational institutions. Additionally, the study proposes recommendations for users of AWE systems and presents a model for effectively incorporating AWE into the learning process.

The findings of this study demonstrated that combined automated-teacher feedback worked as well as teacher feedback when the intervention was over. However, the group in the combined automated-teacher feedback condition improved more from the beginning to the end of the intervention than the group in the instructor feedback condition. While Criterion dealt with lower-level mistakes, the teacher allocated more time to higher-order skills such as content and organization. This situation prevents EFL writing teachers from wasting their energy on correcting repetitive mistakes in grammar, vocabulary, and mechanics, and they can canalize their energy into more meaningful aspects of writing skills.

Furthermore, it is recommended to adopt this combined automated–teacher feedback model so as to enhance the effectiveness of the writing classes, as it contributes to

fostering a positive attitude among students towards writing in English. The results of the study indicated that using an AWE tool served as a source of motivation for students to engage in more revisions and improve their writing, primarily due to automated scoring and the provision of immediate feedback. Moreover, incorporating technology into writing classes has the potential to be engaging and fun for EFL learners.

However, it must be noted that the students and the instructors must be fluent in technology to use AWE writing tools in the classrooms and computers, and a secure internet connection must be available to all parties. This is how maximum efficiency can be received from AWE tools. In addition, the students must be trained in how the system works and how the feedback should be interpreted well, and they must be given time to explore and familiarize themselves with the system.

5.4. Limitations of the Study

There are some limitations to the present study that need to be taken into account when interpreting the results.

The first limitation is the number of students who participated in the study. Although both classes had 26 students on the roster, the participating students were 22 for the experimental group and 16 for the control group.

The second limitation is the duration of the study and the number of essays written. Because this study was conducted at the Basic English Department of the School of Foreign Languages of a state university in Turkey, a strict curriculum related to writing classes needed to be followed. The curriculum allowed only seven weeks to practice opinion essays, and the following seven weeks were allocated for cause-effect essays. As Criterion does not include and rate the cause–effect essay type properly, the research had to stop at that point. During these seven weeks, five essays were written and submitted by the students.

In addition, the research was conducted in an instructional setting, and employing a true experimental research design and having a sample of randomly selected students was not possible. Although the placement test indicated that the students had similar levels of English and went through a similar education in the previous semester, having a true experimental research design would lead to more reliable results.

The final constraint could be associated with the nature of the assignment and the scoring method adopted in this research. The study exclusively utilized persuasive essays written by students, and the results are not generalizable. Moreover, the raters had to employ the holistic rubric created by ETS and implemented by Criterion to evaluate the pre- and post-tests. When using this holistic approach, assessors may rely on their individual opinions to decide which criteria should be accorded more significance during the essay grading process (Barkaoui, 2008; Goulden, 1994).

Finally, due to a devastating earthquake that struck the southeastern part of Turkey on 6th February 2023, Higher Education Council decided that universities should switch to online education during the spring semester. Therefore, this study was conducted online, and if this study had been conducted face-to-face, the participation would have been higher.

5.5. Recommendations for Further Research

Considering the results, limitations, and classroom implications outlined in this study, several recommendations can be proposed for future research. The number of students participating in the study in each group was relatively few. Therefore, more crowded groups can be utilized for research in the future. Additionally, the number of weeks that belonged to intervention and the number of essays written were relatively few. This fact must be considered in future studies.

The following recommendation is about the age of the students. The students in this study were at the tertiary level. However, there are no studies conducted with high school students in the Turkish context, and future studies should take this into account.

Moreover, the correlation between the scores given by the human raters and the scores provided by Criterion can be calculated, and the reliability of the scores given by Criterion can be determined in the context of further studies. These calculations were not done in the scope of this study as Criterion scores were not employed in any test.

Finally, different groups in different feedback conditions can be included in future studies. A group with an automated feedback condition only, a group with a peer feedback condition only, a group with a combined automated – peer feedback condition, etc., maybe a part of future studies. The inclusion of these groups may show

the genuine effect of the use of an AWE tool and combined automated–teacher feedback on the students' writing performance.

REFERENCES / BIBLIOGRAPHY

AbuSeileek, A., & Abualsha'r, A. (2014). Using peer computer-mediated corrective feedback to support EFL learners' writing. *Language Learning & Technology, 18*(1), 76–95.

Allen, L. K, Jacovina, M. E, & McNamara, D. S (2016). Computer-Based Writing Instruction. In C. A., MacArthur, S., Graham, & J. Fitzgerald, (Eds.). (2008), *Handbook of writing research* (pp. 346–361). Guilford Press. https://doi/10.4324/9781410616470

Almusharraf, N. & Hind Alotaibi. (2022). An error-analysis study from an EFL writing context: Human and Automated Essay Scoring Approaches. *Technology, Knowledge, and Learning*. https://doi.org/10.1007/s10758-022-09592-z

Alshakhi, A. (2019). Revisiting the Writing Assessment Process at a Saudi English Language Institute: Problems and Solutions. *English Language Teaching*, *12*(1), 176–185. https://doi/10.5539/elt.v12n1p176

Ary, D., Jacobs, L. C., Irvine, C. K. S., & Walker, D. (2009). *Introduction to research in education* (8th ed). Cengage Learning.

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

Attali, Y. (2013). Validity and Reliability of Automated Essay Scoring. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 181–198). Routledge. https://doi.org/10.4324/9780203122761.

Attali, Y. (2016). A Comparison of Newly-Trained and Experienced Raters on a Standardized Writing Assessment. *Language Testing*, *33*(1), 99–115. https://doi.org/10.1177/0265532215582283

Attali, Y., Bridgeman, B., & Trapani, C. (2010). Performance of a Generic Approach in Automated Essay Scoring. *The Journal of Technology, Learning, and Assessment*, *10*(3), Article 3. https://ejournals.bc.edu/index.php/jtla/article/view/1603

Attali, Y., & Burstein, J. (2006). Automated Essay Scoring With e-rater® V.2. *The Journal of Technology, Learning and Assessment*, *4*(3), Article 3. https://ejournals.bc.edu/index.php/jtla/article/view/1650

Attali, Y., Lewis, W. E., & Steier, M. (2013). Scoring with the computer: Alternative procedures for improving the reliability of holistic essay scoring. *Language Testing*, *30*(1), 125–141. https://doi.org/10.1177/0265532212452396

Attali, Y., & Powers, D. (2009). Validity of scores for a developmental writing scale based on automated scoring. *Educational and psychological measurement*, *69*(6), 978–993.

Bai, L., & Hu, G. (2017). In the face of fallible AWE feedback: How do students respond? *Educational Psychology*, *37*(1), 67–81. https://doi.org/10.1080/01443410.2016.1223275

Barkaoui, K. (2008). *Effects of scoring method and rater experience on ESL essay rating processes and outcomes.* Unpublished doctoral dissertation, University of Toronto, Canada.

Barkaoui, K. (2010a). Variability in ESL Essay Rating Processes: The Role of the Rating Scale and Rater Experience. *Language Assessment Quarterly*, *7*(1), 54–74. https://doi.org/10.1080/15434300903464418

Barkaoui, K. (2010b). Do ESL Essay Raters' Evaluation Criteria Change With Experience? A Mixed-Methods, Cross-Sectional Study. *TESOL Quarterly*, *44*(1), 31–57. https://doi.org/10.5054/tq.2010.214047

Barrot, J. (2021). Using automated written corrective feedback in the writing classrooms: Effects on L2 writing accuracy. *Computer Assisted Language Learning*, pp. 1–24. https://doi.org/10.1080/09588221.2021.1936071

Beigman-Klebanov, B., Burstein, J., Madnani, N., Faulkner, A., & Tetreault, J. (2012). Building subjectivity lexicon(s) from scratch for essay data. In A. Gelbulkh (Ed.), *Springer lecture notes in computer science*. Berlin: Springer-Verlag.

Benali, A. (2021). The Impact of Using Automated Writing Feedback in ESL/EFL Classroom Contexts. *English Language Teaching*, *14*(12), 189–195.

Bereiter, C. (2003). Foreword. In Shermis, M. D., & Burstein, J. C. (Eds). *Automated essay scoring: A cross-disciplinary perspective*. (pp. vii–ix). Routledge. https://doi.org/10.4324/9781410606860

Black, P., & Wiliam, D. (2010). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, *92*(1), 81–90.

Braun, H. I. (1988). Understanding score reliability: Experiments in calibrating essay readers. *Journal of Educational Statistics*, *13*, 1–18. https://doi.org/10.2307/1164948

Bridgeman, B. (2013). Human Ratings and Automated Essay Evaluation. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 221-232). Routledge. https://doi.org/10.4324/9780203122761.

Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education*, *25*(1), 27-40.

Brimi, H. M. (2011). Reliability of Grading High School Work in English. *Practical Assessment, Research, and Evaluation 16* (1), 17. https://doi.org/10.7275/j531-fz38

Burstein, J. (2003). The E-rater® Scoring Engine: Automated Essay Scoring With Natural Language Processing. In Shermis, M. D., & Burstein, J. C. (Eds). *Automated essay scoring: A cross-disciplinary perspective*. (pp. 113–121). Routledge. https://doi.org/10.4324/9781410606860

Burstein, J., Beigman-Klebanov, B., Madnani, N., & Faulkner, A. (2013). Automated Sentiment Analysis for Essay Evaluation. In M. D. Shermis & J. Burstein (Eds.),

Handbook of automated essay evaluation (pp. 281-297). Routledge. https://doi.org/10.4324/9780203122761.

Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The Criterion online writing service. *Ai magazine*, *25*(3), 27-27.

Burstein, J., Kukich, K., Wolff, S., Lu, C., & Chodorow, M. (1998). Computer analysis of essays. *NCME Symposium on Automated Scoring*.

Burstein, J., Tetreault, J., & Madnani, N. (2013). The E-rater® Automated Essay Scoring System. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 55–67). Routledge. https://doi.org/10.4324/9780203122761.

Burstein, J., Tetreault, J., Chodorow, M., Blanchard, D., & Andreyev, S. (2013). Automated Evaluation of Discourse Coherence Quality in Essay Writing. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 267-280). Routledge. https://doi.org/10.4324/9780203122761.

Burston, J. (2001). Computer-Mediated Feedback in Composition Correction. *CALICO Journal*, *19*(1), 37–50.

Chapelle, C. A., Cotos, E., & Lee, J. (2015). Validity Arguments for Diagnostic Assessment Using Automated Writing Evaluation. *Language Testing*, *32*(3), 385–405. https://doi.org/10.1177/0265532214565386

Chen, C.-F. E., & Cheng, W.-Y. (2006). The Use of a Computer-Based Writing Program: Facilitation or Frustration?. *Online Submission*.

Chen, C.-F. E., & Cheng, W.-Y. E. C. (2008). Beyond the Design of Automated Writing Evaluation: Pedagogical Practices and Perceived Learning Effectiveness in EFL Writing Classes. *Language Learning & Technology*, *12*(2), 94–112.

Chen, H. H. J., Chiu, S. T. L., & Liao, P. (2009). Analyzing the Grammar Feedback of Two Automated Writing Evaluation Systems: My Access and Criterion. *English Teaching & Learning*, *33*(2).

Cheng, G. (2017). The impact of online automated feedback on students' reflective journal writing in an EFL course. *The Internet and Higher Education*, *34*, 18–27.

Choi, J. (2010). *The Impact of Automated Essay Scoring (AES) for Improving English Language Learner's Essay Writing* (pp. 1–208). Charlottesville, VA: University of Virginia. https://doi.org/10.18130/V3TC0J

Chukharev-Hudilainen, E., & Saricaoglu, A. (2016). Causal discourse analyzer: Improving automated feedback on academic ESL writing. *Computer Assisted Language Learning*, *29*(3), 494–516.

Chung, G. K. W. K., & O'Neil, H. F. (1997). *Methodological Approaches to Online Scoring of Essays*. https://eric.ed.gov/?id=ED418101

Cohen, J. (1988). Set correlation and contingency tables. *Applied psychological measurement*, *12*(4), 425–434.

Cohen, L., Manion, L., Morrison, K., & Morrison, R. B. (2007). *Research methods in education*. London: Routledge

Conference on College Composition and Communication. (2004). *Position statement on teaching, learning, and assessing writing in digital environments*. Retrieved from http://www.ncte.org/cccc/resources/positions/digitalenvironments.

Conference on College Composition and Communication. (2009). *Writing assessment: A position statement.* Retrieved from http://www.ncte.org/cccc/resources/positions/writingassessment.

Cotos, E. (2015). *Automated Writing Analysis for writing pedagogy: From healthy tension to tangible prospects*.

Creswell, J. W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches.* Sage: Thousand Oakes.

Creswell, J. W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). Boston, MA: Pearson Education.

Creswell, J. W. (2013). *Qualitative inquiry and research design: Choosing among five approaches* (3rd ed.). Sage publications.

Dancey, C. P., & Reidy, J. (2007). Statistics without maths for psychology. Pearson education.

Deane, P. (2013). Covering the Construct: An Approach to Automated Essay Scoring Motivated by a Socio-Conitive Framework for Defining Literacy Skills. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 298–312). Routledge. https://doi.org/10.4324/9780203122761

Dikli, S. (2006). An Overview of Automated Scoring of Essays. *The Journal of Technology, Learning, and Assessment*, *5*(1), Article 1. https://ejournals.bc.edu/index.php/jtla/article/view/1640

Dikli, S. (2010). The Nature of Automated Essay Scoring Feedback. *The CALICO Journal*, *28*(1), 99–134. https://doi.org/10.11139/cj.28.1.99-134

Dikli, S., & Bleyle, S. (2014). Automated Essay Scoring feedback for second language writers: How does it compare to instructor feedback? *Assessing Writing*, *22*, 1–17. https://doi.org/10.1016/j.asw.2014.03.006

Douglas, D. (2009). *Understanding Language Testing*. https://doi.org/10.4324/9780203776339

Ebyary, K. E., & Windeatt, S. (2010). The impact of computer-based feedback on students' written work. *International Journal of English Studies*, *10*(2), Article 2. https://doi.org/10.6018/ijes/2010/2/119231

Elliot, N., & Klobucar A. (2013). Automated Essay Evaluation and the Teaching of Writing. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 16-35). Routledge. https://doi.org/10.4324/9780203122761.

Elliot, S. (2003). Intellimetric: From Here to Validity. In Shermis, M. D., & Burstein, J. C. (Eds). *Automated essay scoring: A cross-disciplinary perspective*. (pp. 71–86). Routledge. https://doi.org/10.4324/9781410606860

Engelhard, G., Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*(2), 93–112. https://doi.org/10.1111/j.1745-3984.1996.tb00483.x

Ericsson, P. F., & Haswell, R. H. (n.d.). *Machine Scoring of Student Essays: Truth and Consequences*.

ETS. (2019). About the CriterionVR Online Writing Evaluation Service. Retrieved from: https://www.ets.org/criterion/about/

Ferris, D. (2011). *Treatment of Error in Second Language Student Writing, Second Edition*. University of Michigan Press.

Foltz, P., Laham, D., & Landauer, T. (1999). The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*.

Foltz, P. W., Streeter, L. A., Lochbaum, K. E. & Landauer T. K. (2013). Implementation and Applications of the Intelligent Essay Assessor. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 68-88). Routledge. https://doi.org/10.4324/9780203122761.

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*(1), 53-80.

Gao, J. (2021). Exploring the Feedback Quality of an Automated Writing Evaluation System Pigai. *International Journal of Emerging Technologies in Learning (Ijet)*, *16*(11), 322–330. https://doi.org/10.3991/ijet.v16i11.19657

Geranpayeh, A. (2003). A quick review of the English Quick Placement Test. *Research notes*, 12(3), 8-10.

Ghufron, M. A., & Rosyida, F. (2018). The Role of Grammarly in Assessing English as a Foreign Language (EFL) Writing. *Lingua Cultura*, *12*(4), 395–403. https://doi.org/10.21512/lc.v12i4.4582

Goulden, N. R. (1994). Relationship of analytic and holistic methods to raters' scores for speeches. *Journal of Research & Development in Education, 27*(2), 73–82.

Grami, G. M. A. (2020). The effects of integrating peer feedback into university-level ESL Writing Curriculum.

Grimes, D. C. (2008). *Middle school use of automated writing evaluation: A multi-site case study* (pp. 1–238). University of California, Irvine.

Grimes, D., & Warschauer, M. (2010). *Utility in a Fallible Tool: A Multi-Site Case Study of Automated Writing Evaluation. 8*(6).

Han, N.R., Chodorow, M., & Leacock, C. (2006). Detecting errors in English article usage by non-native speakers. *Natural Language Engineering*, *12*(2), 115–129. https://doi.org/10.1017/S1351324906004190

Han, T., & Sari, E. (2022). An investigation on the use of automated feedback in Turkish EFL students' writing classes. *Computer Assisted Language Learning*, 1–24.

Hartshorn, K. J., Evans, N. W., Merrill, P. F., Sudweeks, R. R., Strong-Krause, D., & Anderson, N. J. (2010). Effects of Dynamic Corrective Feedback on ESL Writing Accuracy. *TESOL Quarterly*, *44*(1), 84–109. https://doi.org/10.5054/tq.2010.213781

Higgins, D., Burstein, J., & Attali, Y. (2006). Identifying off-topic student essays without topic-specific training data. *Natural Language Engineering*, *12*(2), 145–159.

Hoang, G. T. L., & Kunnan, A. J. (2016). Automated Essay Evaluation for English Language Learners:A Case Study of MY Access. *Language Assessment Quarterly*, *13*(4), 359–376. https://doi.org/10.1080/15434303.2016.1230121

Hockly, N. (2019). Automated writing evaluation.

Hoskins, M., & Wilson, M. (2001). Real-time feedback on rater drift in constructed response items: An example from the Golden State Examination. *Journal of Educational Measurement*, *38*, 121–146. https://doi.org/10.1111/j.1745-3984.2001.tb01119.x

Huang, S. J. (2014). Automated versus human scoring: A case study in an EFL context. *Electronic Journal of Foreign Language Teaching*, *11*, 149–164. https://doi.org/10.1080/15434303.2016.1230121

Huang, J., & Foote, C. J. (2010). Grading between the lines: What really impacts professors' holistic evaluation of ESL graduate student writing? *Language Assessment Quarterly*, *7*(3), 219–233.

Huang, S., & Renandya, W. A. (2020). Exploring the integration of automated feedback among lower-proficiency EFL learners. *Innovation in Language Learning and Teaching*, *14*(1), 15–26. https://doi.org/10.1080/17501229.2018.1471083

Hyland, K., & Hyland, F. (Eds.). (2006). *Feedback in second language writing: Contexts and issues*. Cambridge University Press. https://doi.org/10.1017/CBO9781139524742

Jayavalan, K. & Razali, A. B. (2018). *Effectiveness of Online Grammar Checker to Improve Secondary Students' English Narrative Essay Writing*.

Jingxin, G., & Razali, A. B. (2020). Tapping the potential of Pigai Automated Writing Evaluation (AWE) program to give feedback on EFL writing. *Universal Journal of Educational Research*, *8*(12B), 8334–8343.

Jinla, T. (2014). How to integrate an automated writing assessment tool in the EFL classroom. *Foreign Language Learning Theory and Practice*.

Jones, N. (2013). Reliability and dependability. In *The Routledge Handbook of language testing* (pp. 364–376). Routledge.

Ke, Z., & Ng, V. (2019). Automated Essay Scoring: A Survey of the State of the Art. *International Joint Conference on Artificial Intelligence*, 6300–6308. https://doi.org/10.24963/ijcai.2019/879

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

Kern, R., & Warschauer, M. (2000). Introduction. In M. Warschauer & R. Kern (Eds.), *Network-based language teaching: Concepts and practice* (pp. 1–19). London: Cambridge University Press. https://doi.org/10.1017/CBO9781139524735

Koskey, K. L. K., & Shermis, M. D. (2013). Scaling and Norming for Automated Essay Scoring. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated*

essay evaluation (pp. 199-220). Routledge. https://doi.org/10.4324/9780203122761.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2–3), 259–284. https://doi.org/10.1080/01638539809545028

Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated Scoring and Annotation of Essays with the Intelligent Essay Assessor™. In Shermis, M. D., & Burstein, J. C. (Eds). *Automated essay scoring: A cross-disciplinary perspective*. (pp. 87-112). Routledge. https://doi.org/10.4324/9781410606860

Latifi, S. M. F. (2016). *Development and validation of an automated essay scoring framework by integrating deep features of English language* [Unpublished doctoral dissertation]. University of Alberta, Edmonton, Canada. https://era.library.ualberta.ca

Lavolette, E., Polio, C., & Kahng, J. (2015). The accuracy of computer-assisted feedback and students' responses to it. *Language, Learning & Technology*, *19*(2).

Lee, Y. J. (2020). The Long-Term Effect of Automated Writing Evaluation Feedback on Writing Development. *English Teaching*, *75*(1), 67–92. https://doi.org/10.15858/engtea.75.1.202003.67

Li, J., Link, S., & Hegelheimer, V. (2015). Rethinking the role of automated writing evaluation (AWE) feedback in ESL writing instruction. *Journal of Second Language Writing*, *27*, 1–18. https://doi.org/10.1016/j.jslw.2014.10.004

Li, Z., Zhi Li, Link, S., Ma, H., Yang, H., & Hegelheimer, V. (2014). The role of automated writing evaluation holistic scores in the ESL classroom. *System*, *44*, 66–78. https://doi.org/10.1016/j.system.2014.02.007

Liao, H. C. (2016). Using automated writing evaluation to reduce grammar errors in writing. *Elt Journal*, *70*(3), 308–319.

Liddy, E. D. (2001). Natural language processing. In *Encyclopedia of library and information science* (2 ed.). New York: Marcel Decker, Inc.

Lim, C. T., Bong, C. H., Wong, W. S., & Lee, N. K. (2021). A comprehensive review of automated essay scoring (AES) research and development. *Pertanika Journal of Science & Technology*, *29*(3), 1875–1899.

Link, S., Mehrzad, M., & Rahimi, M. (2020). Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement. *Computer Assisted Language Learning*, 1–30. https://doi.org/10.1080/09588221.2020.1743323

Lottridge, S. M., Schulz, E. M., & Mitzel, H. C. (2013). Using Automated Scoring to Monitor Reader Performance and Detect Reader Drift in Essay Scoring. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 233-250). Routledge. https://doi.org/10.4324/9780203122761.

Lv, X. (2018). A Study on the Application of Automatic Scoring and Feedback System in College English Writing. *International Journal of Emerging Technologies in Learning (Ijet)*, *13*(3), 188–196. https://doi.org/10.3991/ijet.v13i03.8386

McCurry, D. (2010). Can Machine Scoring Deal with Broad and Open Writing Tests as Well as Human Readers. *Assessing Writing*, *15*(2), 118–129. https://doi.org/10.1016/j.asw.2010.04.002

Measurement Incorporated. (2020). *Automated essay scoring*. Retrieved January 13, 2023, from https://www.measurementinc.com/products-services/automated-essay-scoring

Mehrabi-Yazdi, O. (2018). Short communication on the missing dialogic aspect of an automated writing evaluation system in written feedback research. *Journal of Second Language Writing*, *41*, 92-97.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Milton, J. (2006). Resource-rich Web-based feedback: Helping learners become independent writers. In K. Hyland & F. Hyland (Eds.), *Feedback in second language writing: Context and issues* (pp. 123–139). New York: Cambridge University Press. https://doi.org/10.1017/CBO9781139524742

Mohsen, M. A., & Alshahrani, A. (2019). The Effectiveness of Using a Hybrid Mode of Automated Writing Evaluation System on EFL Students' Writing. *Teaching English with Technology*, *19*(1), 118–131.

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, *18*(5), 544–551.

National Research Council. (2001). *Classroom assessment and the National Science Education Standards.* Washington, DC: National Academies Press. Retrieved from http://www.nap.edu/catalog/9847/classroom-assessment-and-the-national-science-education-standards.

Page, E. B. (1966). The imminence of... grading essays by computer. *The Phi Delta Kappan*, *47*(5), 238–243.

Page, E. B. (1994). Computer grading of student prose using modern concepts and software. *The Journal of experimental education*, *62*(2), 127–142. https://doi.org/10.1080/00220973.1994.9943835

Page, E. B. (2003). Project Essay Grade: PEG. In Shermis, M. D., & Burstein, J. C. (Eds). *Automated essay scoring: A cross-disciplinary perspective*. (pp. 443–54). Routledge. https://doi.org/10.4324/9781410606860

Page, E. B., & Petersen, N. S. (1995). The computer moves into essay grading: Updating the ancient test. *Phi Delta Kappan*, *76*(7), 561.

Page, E. B., Poggio, J. P., & Keith, T. Z. (1997). *Computer Analysis of Student Essays: Finding Trait Differences in Student Profile.*

Pallant, J. (2020). *SPSS survival manual: A step by step guide to data analysis using IBM SPSS.*

Park, J. (2019). An AI-based English Grammar Checker vs. Human Raters in Evaluating EFL Learners' Writing. *Multimedia-Assisted Language Learning*, *22*(1), 112–131.

Parra G, L., & Calero S, X. (2019). Automated writing evaluation tools in the improvement of the writing skill. *International Journal of Instruction*, *12*(2), 209-226.

Plano Clark, V. L., & Creswell, J. W. (2015). *Understanding research: A consumer's guide*. Pearson.

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2001). Stumping E-Rater: Challenging the validity of automated essay scoring. *ETS Research Report Series*, *2001*(1), i-44.

Quellmalz, E. (1980). *Problems in stabilizing the judgment process* (Vol. CSE Report No. 136). Los Angeles, CA: Center for the Study of Evaluation.

Ranalli, J. (2018). Automated written corrective feedback: How well can students make use of it? *Computer Assisted Language Learning*, *31*(7), 653–674. https://doi.org/10.1080/09588221.2018.1428994

Ranalli, J., Link, S., & Chukharev-Hudilainen, E.. (2017). Automated writing evaluation for formative assessment of second language writing: investigating the accuracy and usefulness of feedback as part of argument-based validation. *Educational Psychology*, *37*(1), 8–25. https://doi.org/10.1080/01443410.2015.1136407

Rao, Z., & Li, X. (2017). Native and Non-native Teachers' Perceptions of Error Gravity: The Effects of Cultural and Educational Factors. *Asia-Pacific Education Researcher*, *26*(1), 51–59. https://doi.org/10.1007/s40299-017-0326-5

Ravand, H., & Rasekh, A. E. (2011). Feedback in ESL Writing: Toward an Interactional Approach. *Journal of Language Teaching and Research*, *2*(5), 1136–1145. https://doi.org/10.4304/jltr.2.5.1136-1145

Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing*, *15*(1), 18–39. https://doi.org/10.1016/j.asw.2010.01.003

Rock, J. L. (2007). THE IMPACT OF SHORT-TERM USE OF CRITERION ON WRITING SKILLS IN NINTH GRADE. *ETS Research Report Series*, *2007*(1), i-24.

Rose, D. (2016). New Developments in Genre-Based Literacy Pedagogy. In C. A., MacArthur, S., Graham, & J. Fitzgerald, (Eds.). (2008), *Handbook of writing research* (pp. 250–269). Guilford Press. https://doi/10.4324/9781410616470

Rudner, L. M., & Gagne, P. (2000). An overview of three approaches to scoring written essays by computer. *Practical Assessment, Research, and Evaluation*, *7*(1), 26. https://doi/10.47836/pjst.29.3.27.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, *88*(2), 413–428. https://doi.org/10.1037/0033-2909.88.2.413

Scharber, C., Dexter, S., & Riedel, E. (2008). Students' Experiences with an Automated Essay Scorer. *Journal of Technology, Learning, and Assessment*, *7*(1), n1.

Sharma, C., Bishnoi, A., Sachan, A. K., & Verma, A. (2019). Automated essay evaluation using natural language processing. *International Research Journal of Engineering and Technology*, *5*(6), 2055–2058. https://www.irjet.net/archives/V6/i5/IRJET-V6I5398.pdf

Schroeder, J., Grohe, B., & Pogue, R. (2008). The impact of criterion writing evaluation technology on criminal justice student writing skills. *Journal of Criminal Justice Education, 19*(3), 432-445.

Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing*, *20*, 53–76.

Shermis, M. D., & Burstein, J. C. (2003). *Automated essay scoring: A cross-disciplinary perspective*. Routledge. https://doi.org/10.4324/9781410606860

Shermis, M. D., & Burstein, J. C. (2003). Introduction. In Shermis, M. D., & Burstein, J. C. (Eds). *Automated essay scoring: A cross-disciplinary perspective*. (pp. xiii -xvi). Routledge. https://doi.org/10.4324/9781410606860

Shermis, M. D., & Burstein, J. (2013). Handbook of automated essay evaluation. *NY: Routledge*.

Shermis, M. D., Burstein, J., & Bursky, S. A. (2013). Introduction to automated essay evaluation. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 1-15). Routledge. https://doi.org/10.4324/9780203122761.

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

Shin, J., & Gierl, M. J. (2021). More efficient processes for creating automated essay scoring frameworks: A demonstration of two algorithms. *Language Testing*, *38*(2), 247–272. https://doi.org/10.1177/0265532220937830

Stevenson, M. (2016). A Critical Interpretative Synthesis: The Integration of Automated Writing Evaluation into Classroom Writing Instruction. *Computers and Composition*, *42*, 1–16. https://doi.org/10.1016/j.compcom.2016.05.001

Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. *Assessing Writing*, *19*, 51–65. https://doi.org/10.1016/j.asw.2013.11.007

Stiggins, R. J., & Bridgeford, N. J. (1983). An analysis of published tests of writing proficiency. *Educational Measurement: Issues and Practice*, *2*(1), 6–6.

Tang, J., & Rich, C. S. (2017). Automated Writing Evaluation in an EFL Setting: Lessons from China. *JALT CALL Journal*, *13*(2), 117–146.

Tetreault, J. R., & Chodorow, M. (2008). The Ups and Downs of Preposition Error Detection in ESL Writing. *Proceedings of the 22nd International Conference on Computational Linguistics*, 865–872. https://aclanthology.org/C08-1109

Tolchinsky, L. (2016). From Text to Language and Back: The Emergence of Written Language. In C. A., MacArthur, S., Graham, & J. Fitzgerald, (Eds.). (2008), *Handbook of writing research* (pp. 156-173). Guilford Press. https://doi/10.4324/9781410616470

Tömen, M. (2022). Automated Essay Scoring Feedback in Foreign Language Writing: Does it Coincide with Instructor Feedback? *Disiplinler Arası Dil Araştırmaları*, *4*(4), 53–62.

Vantage Learning (2001). Applying IntelliMetric™ to the scoring of entrylevel college student essays. (RB-539). Retrieved April 6, 2023, from https://www.vantage.com/pdfs/research/RB539.pdf

Wali, F. & Huijser, H. (2018). Write to improve: Exploring the impact of an automated feedback tool on Bahraini learners of English. *Learning & Teaching in Higher Education: Gulf Perspectives*, *15*(1), 14–34. https://doi.org/10.18538/lthe.v15.n1.293

Wang, J., & Brown, M. S. (2007). Automated essay scoring versus human scoring: A comparative study. *Journal of Technology, Learning, and Assessment*, *6*(2), n2.

Wang, P. (2013). Can Automated Writing Evaluation Programs Help Students Improve Their English Writing?. *International Journal of Applied Linguistics and English Literature*, *2*(1), 6–12. https://doi.org/10.7575/ijalel.v.2n.1p.6

Wang, Y.-J., Shang, H.-F., & Briody, P. (2013). Exploring the impact of using automated writing evaluation in English as a foreign language university students' writing. *Computer Assisted Language Learning*, *26*(3), 234–257. https://doi.org/10.1080/09588221.2012.655300

Warschauer, M., & Grimes, D. (2008). Automated Writing Assessment in the Classroom. *Pedagogies: An International Journal*, *3*(1), 22–36. https://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. https://doi.org/10.1191/1362168806lr190oa

Weigle, S. C. (2002). *Assessing Writing*. Cambridge University Press.

Weigle, S. C. (2012). Assessing writing. In Peter Davidson et al. (Eds.), *The Cambridge guide to language assessment.* Cambridge: Cambridge University Press

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

Wilson, J., Olinghouse, N. G., McCoach, D. B., Santangelo, T., & Andrada, G. N. (2016). Comparing the accuracy of different scoring methods for identifying sixth graders at risk of failing a state writing assessment. *Assessing Writing*, *27*, 11–23.

Wilson, J., & Roscoe, R. D. (2020). Automated Writing Evaluation and Feedback: Multiple Metrics of Efficacy: *Journal of Educational Computing Research*, *58*(1), 87–125. https://doi.org/10.1177/0735633119830764

Wind, S. A. & Engelhard, G. (2013). How invariant and accurate are domain ratings in writing assessment?. *Assessing Writing*, *18*(4), 278–299. https://doi.org/10.1016/j.asw.2013.09.002

Wiseman, C. S. (2012). Rater effects: Ego engagement in rater decision-making. *Assessing Writing*, *17*(3), 150–173. https://doi.org/10.1016/j.asw.2011.12.001

Wolfe, M. B., Schreiner, M. E., Rehder, B., Laham, D., Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). Learning from text: Matching readers and texts by latent semantic analysis. *Discourse Processes*, *25*(2-3), 309-336.

Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). Second Language Development in Writing: Measures of Fluency, Accuracy, and Complexity. University of Hawaii Press.

Yeh, Y. L., Liou, H. C., & Yu, Y. T. (2007). The influence of automatic essay evaluation and bilingual concordancing on EFL students' writing. *英語教學期刊*, *31*(1), 117-160.

Zaidi, A. H. (2016). *Neural sequence modelling for automated essay scoring* [Unpublished MA thesis]. University of Cambridge. https://www.cl.cam.ac.uk/

Zhang, M. (2013). Contrasting automated and human scoring of essays. *R & D Connections*, *21*(2). https://www.ets.org/Media/Research/pdf/RD_Connections_21.pdf

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

Zhang, Z., & Hyland, K. (2018). Student engagement with teacher and automated feedback on L2 writing. *Assessing Writing*, *36*, 90–102. https://doi.org/10.1016/j.asw.2018.02.004.

APPENDICES

Appendix 1. Criterion Rubric

ETS Criterion Score Guide

6 | 5 | 4 | 3 | 2 | 1

Score = 6 — **A typical essay at this level:**

- effectively addresses the writing task
- is well organized and well developed
- uses clearly appropriate details to support a thesis or illustrate ideas
- displays consistent facility in the use of language
- demonstrates syntactic variety and appropriate word choice, though it may have occasional errors

Score = 5 — **A typical essay at this level:**

- may address some parts of the task more effectively than others
- is generally well-organized and well-developed
- uses details to support a thesis or illustrate idea
- displays facility in the use of language
- demonstrates some syntactic variety and range of vocabulary, though it will probably have occasional errors

Score = 4 — **A typical essay at this level:**

- addresses the writing topic adequately but may slight parts of the task
- is adequately organized and developed
- uses some details to support a thesis or illustrate an idea
- demonstrates adequate but possibly inconsistent facility with syntax and usage
- may contain some errors that occasionally obscure meaning

Score = 3 — **A typical essay at this level may reveal one or more or the following weaknesses:**

- inadequate organization or development
- inappropriate or insufficient details to support or illustrate generalizations
- a noticeably inappropriate choice of words or word forms
- an accumulation of errors in sentence structure and/or usage

Score = 2 — **A typical essay at this level is flawed by one or more of the following weaknesses:**

- serious disorganization or underdevelopment
- little or no detail, or irrelevant specifics
- serious and frequent errors in sentence structure and usage
- serious problems with focus

Score = 1 — **A typical essay at this level:**

- may be incoherent
- may be undeveloped
- may contain severe or persistent writing errors

Appendix 2. One-on-one Interview Questions

Q1. What do you think about the scores given by Criterion? To what extent did they reflect your writing performance?

Q2. What did you do to increase the score that you got from Criterion?

Q3. What do you think about the feedback you got from Criterion? How was it different from the teacher feedback? Which one do you prefer – Criterion feedback, teacher feedback, or combined automated-teacher feedback?

Q4. To what extent do you think using Criterion improved your writing skills? In which area do you think you improved the most?

Q5. What are the strengths and weaknesses of Criterion?

Q6. How did using Criterion affect your motivation?

Q7. Do you have any suggestions related to the whole process or the use of Criterion?

Appendix 3. Ethics Committee Approval



YILDIZ TEKNİK ÜNİVERSİTESİ
Sosyal ve Beşeri Bilimler Araştırmaları Etik Kurulu

Toplantı Tarihi: 28.12.2022 **Toplantı No: 2022.12**

SOSYAL VE BEŞERİ BİLİMLER ARAŞTIRMALARI ETİK KURULU TOPLANTI KARARI

Yürütücülüğünü Üniversitemiz YABANCI DİLLER YÜKSEKOKULU öğretim üyelerinden Doç. Dr. Suzan Kavanoz danışmanlığında lisansüstü öğrencisi Merve BAKO tarafından yapılacak olan "Otomatik Kompozisyon Değerlendirme Araçlarının İngilizce'Yi Yabancı Dil Olarak Öğrenen Öğrencilerin Yazma Performansı Üzerindeki Etkisi" adlı çalışma ve bu çalışmada kullanılacak veri toplama araçlan ve yöntemlerine ilişkin bilgilerde etiğe aykın herhangi bir bulguya rastlanmamıştır.

Etik Kurul Üyeleri

Prof. Dr. Murat DÖNDÜREN
Başkan

Prof. Dr. Ali ERYILMAZ
Üye

Prof. Dr. Gülnayat GOLBAŞI ŞIMŞEK
Üye

Doç. Dr. Mehmet Emin KAHRAMAN
Üye

Doç. Dr. Senay OĞUZTİMUR
Üye

Doç. Dr. Yasin ŞEHİTOĞLU
Üye

Dr. Öğr. Üyesi Güzin AKYILDIZ
Üye

Taş Bina A-2013 Oda Davutpaşa Kampüsü Esenler/İstanbul Tel: (0212) 383 31 18
E-posta: sbemdr@yildiz.edu.tr - Web: sbe.yildiz.edu.tr
Rapor Doğrulama Adresi: "etik.yildiz.edu.tr/dogrula" Rapor No: 20221201821 Doğrulama Kodu: 377eb

Appendix 4. Informed Consent Form 1

DENEKLERİN GÖNÜLLÜLÜĞÜ

VE

AYDINLATILMIŞ ONAM FORMU

Sayın Katılımcı,

Katıldığınız bu çalışma bilimsel nitelikte bir araştırmadır. Araştırmanın konusu "Otomatik Kompozisyon Değerlendirme Araçlarının İngilizce'yi Yabancı Dil Olarak Öğrenen Öğrencilerin Yazma Performansı Üzerindeki Etkisi"dir. Bu araştırma, Yıldız Teknik Üniversitesi Sosyal Bilimler Enstitüsü Yabancı Diller Eğitimi Ana Bilim Dalı'nda akademik eğitimini sürdüren Merve Bako'nun yüksek lisans tezi kapsamında yapılmaktadır.

Bu araştırmanın amacı, otomatik kompozisyon değerlendirme araçlarını kullanan öğrencilerin kullanmayan öğrencilere göre yazma becerilerinin daha fazla gelişme gösterip göstermediğini görmektir. Araştırma boyunca sizlerden kişisel bilgileriniz istenmeyecektir. Verdiğiniz yanıtlardan ve tamamladığınız ölçeklerden elde edilecek veriler yalnızca araştırmacı tarafından, araştırma kapsamı dahilinde kullanılacaktır. Bu araştırmaya katılmanızda öncül esas gönüllü olmanızdır. Çalışmaya katılmama ve çalışmanın herhangi bir aşamasında, hiçbir yaptırıma maruz kalmaksızın çalışmadan ayrılma hakkına sahipsiniz.

Yukarıdaki ifadeleri okuduktan sonra çalışmaya katılmak istiyorsanız, aşağıda verilen yerleri tamamlayınız ve imza atınız.

Yukarıda yer alan bilgiler bana araştırmacı tarafından sözlü ve yazılı olarak bildirilmiştir. Çalışmaya katılmayı hiçbir baskı altında kalmadan, kendi rızamla kabul ediyorum.

Katılımcının;

Adı-Soyadı:

Tarih:

İmza:

Appendix 5. Informed Consent Form 2

BİLGİLENDİRİLMİŞ OLUR FORMU

1. Aşağıda imzası olan ben " Otomatik Kompozisyon Değerlendirme Araçlarının İngilizce'yi Yabancı Dil Olarak Öğrenen Öğrencilerin Yazma Performansı Üzerindeki Etkisi" başlıklı

çalışmaya katılmayı kabul ediyorum.

2. Bu çalışmayı yürüten Merve Bako çalışmanın yapısı, amacı ve muhtemel süresi, hakkında ayrıntılı sözlü ve yazılı bilgi verdi.

3. Araştırmacı Merve Bako'ya çalışmasıyla ilgili her soruyu sorma fırsatını buldum. Cevapları ve bana verilen bilgiyi anladım.

4. Araştırmacı Merve Bako'ya bilgilerin ayrıntılarını açıklamama ve benimle ilgili sırları koruması şartıyla benimle bu çalışmayı yapmasına izin veriyorum.

5. Çalışma boyunca tüm kurallara uyacağıma, araştırmacı Merve Bako ile tam bir uyum içinde çalışacağıma ve konuyla ilgili herhangi bir sorun çıktığında hemen onu arayacağımı kabul ediyorum.

6. Bu çalışma sonuçlarının kullanılmasını kısıtlamayacağımı ve yayın, rapor ve benzeri bilimsel dokümanlarda kullanabileceğini kabul ediyorum.

7. Bu çalışmadan istediğim zaman çıkabileceğimi anladım.

OKUDUM VE ONAYLADIM.

Katılımcının Adı ve Soyadı, Adresi: Tarih, İmza:

Araştırmacının Adı ve Soyadı, Adresi: Tarih, İmza:

Appendix 6. Error Correction Code

WRITING ERROR CORRECTION CODE

Symbol	Meaning of Error	Example with Error	Corrected Version
S/V	subject/verb agreement error	She talk too much. S/V	She ***talks*** too much.
S/P	singular/plural error	They have three child. S/P	They have three ***children***.
S	spelling error	Her children go to scool. S	Her children go to ***school***.
↻	move this word	She slowly talks.	She talks ***slowly***.
A	article error	There is the ball in the garden. A	There is ***a*** ball in the garden.
∧	add a word/missing word	She to school. ∧	She ***goes*** to school.
P	punctuation error	Do they have a house on this street. P	Do they have a house on this street***?***
W.O.	word order	She to the park went. W.O.	She ***went to the park***.
W.F.	word form	This film is very bored. W.F.	This film is very ***boring***.
Prep.	preposition error	Classes start on 8:40. Prep. Classes start 8:40.	Classes start ***at*** 8:40. Classes start ***at*** 8:40.
T	incorrect tense	Yesterday, she see her boyfriend. T	Yesterday, she ***saw*** her boyfriend.
WW	wrong word	She has yellow hair. WW	She has ***blonde*** hair.
✕	remove this word(s)	I have to study my lesson for the test.	I have to study for the test.
RW	rewrite	The man knows is old. RW	The man ***that he knows*** is old.
⌣	join together	They went to Paris. They went to London. They went to Paris al together.	They went to Paris and London. They went to Paris ***altogether***.
⊏	separate	They looked at eachother. ⊏	They looked at ***each other***.
Ref.	unclear reference	We saw some people. He was happy. Ref.	We saw some people. ***They*** were happy.
Aux.	auxiliary verb (am/is/are/was/were/modals) error	Men be always different from women. Aux.	Men ***are*** always different from women.
Inc. Sen.	incomplete sentence	Because I am hungry. Inc. Sen.	***I want to buy a burger*** because I am hungry.
Inf.	informal language	Susan has two kids. Inf.	Susan has two ***children***.
Run-on	run-on sentence	My car is broken we cannot reach town before dark. Run-on	a) My car is broken ***so*** we cannot reach town before dark. b) My car is broken**.** We cannot reach town before dark.
Paral.	parallelism error	I like reading, swimming, and to cycle. Paral.	I like reading, swimming, and ***cycling.***
?	not clear / what do you mean?		