

Exploring the Impact of Automated Written Feedback on EFL Students Performance in Essay Writing

Nadia Ahmed Junaid

PhD in TESOL

Department of Education, University of York

March 2025

Abstract

Automated written feedback has emerged as a valuable tool in language learning, offering learners opportunities for independent revision and immediate error correction. Its role in developing EFL writing skills has gained increasing attention, particularly when used alongside traditional teacher feedback. This study investigates the effectiveness of different modes of written feedback—teacher-only, automated-only, and hybrid (automated + teacher)—on the essay writing quality of EFL learners in a Saudi higher education context. It also explores whether the type of writing genre (expository or persuasive) interacts with feedback type to influence student performance.

To examine these issues, a quasi-experimental design was adopted. Seventy-four foundation-year students at a Saudi university participated in the study. All participants were intermediate English learners. Students were divided into three feedback groups and completed writing tests at three stages: pretest, immediate posttest, and delayed posttest. Essay quality was evaluated across five components: text content, text organization, vocabulary use, language use, and mechanics. The writing tasks included both persuasive and expository essays, and scores were analyzed using mixed-effects modeling to assess changes over time and across feedback conditions.

The results showed that all three feedback types contributed to short-term improvements in writing quality. At the delayed posttest, the hybrid feedback group demonstrated the greatest gains in scores. No significant differences were found in the impact of feedback condition when the writing genre changed. Nonetheless, participants across all groups showed greater improvement when writing the more complex genre (persuasive). Overall, the findings suggest

that integrating automated feedback with teacher guidance may support long-term writing development more effectively than using either approach alone.

Table of Contents

ABSTRACT 2
TABLE OF CONTENTS 4
ACKNOWLEDGEMENT 7
DECLARATION 8
LIST OF APPENDICES 9
LIST OF TABLES 10
LIST OF FIGURES 12
1. INTRODUCTION 13
1.1 RATIONALE AND BACKGROUND 13
1.2 CONTEXT OF THE STUDY 14
1.3 THE USE OF TECHNOLOGY IN THE EFL CLASSROOM AND THE VISION FOR SAUDI ARABIA . 17
1.4 PERSONAL MOTIVATION 18
1.5 PURPOSE AND RESEARCH QUESTIONS 20
1.6 STRUCTURE OF THE THESIS 23
2. LITERATURE REVIEW: THEORETICAL FOUNDATION 25
2.1 OVERVIEW 25
2.2 FEEDBACK ON WRITING: AN INTRODUCTION 25
2.3 IMPORTANCE OF INTERACTION 29
2.4 INPUT, OUTPUT AND NOTICING 34
2.5 TASK COMPLEXITY 36
2.6 SOCIOCULTURAL THEORY 39
2.6.1 Mediation in Language Learning 40
2.1.1 Zone of Proximal Development 43
2.2 THE ROLE OF TECHNOLOGY IN SECOND LANGUAGE WRITING 46
2.3 SUMMARY 51
3. EMPIRICAL RESEARCH ON THE IMPACT OF AWE ON WRITING QUALITY 53
3.1 OVERVIEW 53
3.2 AUTOMATED WRITING EVALUATION DEFINITION AND SIGNIFICANCE 53
3.3 AWE TOOLS: FUNCTIONS AND FEEDBACK MECHANISMS 54
3.4 THE EFFECTIVENESS OF AWE AS A SCORING SYSTEM 59

3.5 The Impact of AWE Feedback on Writing Quality 61
3.6 Challenges and Advantages in AWE Implementation 70
3.7 AWE Feedback vs. Teacher Written Feedback 72
3.8 AWE and Teacher Feedback: A Combined Approach 74
3.9 Summary 79
4. Methodology 83
4.1 Overview 83
4.2 Research Questions 83
4.3 Research Design 83
4.4 Participants and Setting 85
4.4.1 Students 85
4.4.2 Teachers 88
4.5 Data Collection 89
4.5.1 Instruments 89
4.5.2 Scoring 94
4.5.3 Treatment Sessions 95
4.5.4 Procedure 96
4.5.5 Criterion Software 103
4.6 Data Analysis 104
4.6.1 Operationalisation of Measures 106
4.6.2 One-way ANOVA 109
4.6.3 Descriptive Statistics 109
4.6.4 Inferential Statistics 110
4.6.5 Reporting Results 112
4.7 Pilot Study 116
4.8 Validity and Reliability 117
4.9 Ethical Considerations 119
4.10 Summary 121
5. Results 122
5.1 Overview 122
5.2 Research Question 1: L2 Overall Writing Production Predicted Relative to

THE FEEDBACK CONDITION 123
5.3 RESEARCH QUESTION 2: COMPARISONS OF THE EFFECTS OF FEEDBACK CONDITION ON L2 WRITING COMPONENTS (TEXT CONTENT, TEXT ORGANISATION, VOCABULARY USE, LANGUAGE USE AND MECHANICS) 130
5.3.1 Text Content 130
5.3.2 Text Organisation 136
5.3.3 Vocabulary Use 142
5.3.4 Language Use 148
5.3.5 Mechanics 154
5.4 RESEARCH QUESTION 3: COMPARISON OF THE EFFECTS OF FEEDBACK CONDITION AND WRITING GENRE (EXPOSITORY, PERSUASIVE) ON OVERALL L2 WRITING PRODUCTION 161
5.4.1 Teacher Feedback Group 163
5.4.2 Automated Feedback Group 165
5.4.3 Hybrid (Automated+teacher) Feedback Group 166
5.5 SUMMARY OF THE RESULTS 170
6. DISCUSSION 172
6.1 OVERVIEW 172
6.2 KEY FINDINGS 172
6.3 THE IMPACT OF FEEDBACK TYPE ON L2 STUDENTS WRITING PERFORMANCE 174
6.4 THE IMPACT OF GENRE TYPE AND FEEDBACK CONDITION ON L2 WRITING PERFORMANCE 181
6.7 SUMMARY 183
7. CONCLUSION 184
7.1 OVERVIEW 184
7.2 SUMMARY OF MAIN FINDINGS 185
7.3 STUDY CONTRIBUTIONS 187
7.3 LIMITATIONS 194
7.4 FUTURE RESEARCH DIRECTIONS 195
APPENDICES 198
REFERENCES 243

Acknowledgement

First and foremost, I would like to express my deepest gratitude to my supervisor, Dr Nadia Mifka-Profozic, for her invaluable guidance, constant encouragement, and unwavering support throughout this research journey. I also extend my thanks to Dr. David O'Reilly whose input and constructive comments were greatly appreciated during the key stages of this thesis.

The completion of this thesis would not have been possible without the love and support of my family. To my father and mother, your unconditional support, help, and prayers gave me the strength to keep going—I am forever grateful.

To my husband, thank you for your endless patience, encouragement, and for standing beside me during every challenge. Your support made it possible for me to keep going. To my three wonderful children, you were my motivation through the hardest days. I hope this journey inspires you to chase your own dreams with determination.

I am also thankful to everyone who contributed to this study. My heartfelt appreciation goes to Ashjan, Ebtehal, Noura, Maha, Elham, and Amani—thank you for welcoming me into your classrooms and encouraging your students to participate in my research. I also extend my thanks to all the students who took part in this study.

To my dear friend Hajar, thank you for being there through the ups and downs of this PhD journey. Your friendship made this experience easier and more meaningful.

This work was supported by University of Jeddah and the Saudi Arabian Cultural Bureau, to them I owe gratitude for their continuous help and generous provisions throughout my scholarship.

Declaration

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for a degree or other qualification at this University or elsewhere. All sources are acknowledged as references.

List of Appendices

Appendix A: Analytic Rubric Adapted from Connor-Linton & Polio (2014)……198

Appendix B: List of Categories and Subcategories of Criterion Feedback.....200

Appendix C: Pre-, Post-, and Delayed Test Prompts 203

Appendix D: Treatment Sessions Task Prompts.....206

Appendix E: QQ Plots for Overall Writing Production 207

Appendix F: Histograms for Overall Writing Production.....208

Appendix G: Mixed-Effects Model Outcomes 210

Appendix H: Consent Form 231

List of Tables

Table 1 A Proposed Plan for Integrating AWE into EFL writing instruction. 57
Table 2 Tests Counterbalancing.....93
Table 3 Single Score Intraclass Correlation Coefficient (ICC). 95
Table 4 Explanation of Variables.....107
Table 5 Summary of Fixed Effects Estimates (Reference: Teacher Feedback Group and Pretest): Example for Illustration. 115
Table 6 Descriptive Statistics for Overall Writing Scores in the Pretest, Immediate Posttest and Delayed Posttest.....124
Table 7 Within-Group Comparisons of Mixed-Effects Model Outcomes for Overall Writing Production. 126
Table 8 Between-Group Comparisons of Mixed-Effects Model Outcomes for Overall Writing Production.....128
Table 9 Summary of Descriptive Statistics for Text Content Scores Obtained at Pretest, Immediate Posttest and Delayed Posttest. 131
Table 10 Within-Group Comparisons of Mixed-effects Model Outcome for Text Content Scores.....132
Table 11 Between-Group Comparison of Mixed-effects Model Outcome for Text Content Scores. 133
Table 12 Summary of Descriptive Statistics for Text Organisation Scores at Pretest, Immediate Posttest and Delayed Posttest.....136
Table 13 Within-Group Comparisons of Mixed-effects Model Outcome for Text Organisation Scores. 138
Table 14 Between-Group Comparison of Mixed-effects Model Outcome for Text Organisation Scores. 140
Table 15 Summary of Descriptive Statistics for Vocabulary Use Scores at Pretest, Immediate Posttest and Delayed Posttest.....142
Table 16 Within-Group Comparisons of Mixed-effects Model Outcome for Vocabulary Use Scores.....144
Table 17 Between-Group Comparisons of Mixed-effects Model Outcome for Vocabulary Use Scores. 146
Table 18 Summary of Descriptive Statistics for Language Use Scores Recorded in Pretest, Immediate Posttest and Delayed Posttest. 148
Table 19 Within-Group Comparisons of Mixed-effects Model Outcome for Language Use Scores.....150
Table 20 Between-Group Comparisons of Mixed-effects Model Outcomes for Language Use Scores. .152
Table 21 Summary of Descriptive Statistics for Mechanics Scores Recorded in Pretest, Immediate

Posttest and Delayed Posttest. 154
Table 22 Within-Group Comparisons of Mixed-effects Model Outcomes for Mechanics Scores. 156
Table 23 Between-Group Comparisons of Mixed-effects Model Outcomes for Mechanics Scores. 158
Table 24 Summary of Descriptive Statistics for Writing Performance in the Expository Genre at Pretest, Immediate Posttest and Delayed Posttest. 162
Table 25 Summary of Descriptive Statistics for Writing Performance in the Persuasive Genre at Pretest, Immediate Posttest and Delayed Posttest for the Three Feedback Groups. 163
Table 26 Teacher Feedback Group: Results of the Linear Mixed-effects Model for the Writing Genres Examined (Expository, Persuasive). 164
Table 27 Automated Feedback Group: Results of the Linear Mixed-effects Model for the Writing Genres Examined. 166
Table 28 Hybrid feedback group: results of linear mixed-effects model for the writing genre examined. 167

List of Figures

Figure 1 Mediation in Sociocultural Theory. 41
Figure 2 A Flowchart of Groups Assignment Process. 87
Figure 3 Example of Teacher Feedback Inserted in Criterion for the Participants (Hybrid Feedback Group). 101
Figure 4 A diagram of the Study Design. 102
Figure 5 Group Overall Performance in L2 Writing Over Time. 125
Figure 6 Plot of the Linear Mixed-effects Model Outcome of Total Writing Scores. 127
Figure 7 Plot of the Linear Mixed-effects Model Outcome for Text Content Scores. 134
Figure 8 Plot of the Linear Mixed-effects Model Outcome for Text Organisation Scores. 139
Figure 9 Plot of the Linear Mixed-effects Model Outcome of Vocabulary Use Scores. 145
Figure 10 Plot of the Linear Mixed-effects Model Outcome of Language Use Scores. 151
Figure 11 Plot of the Linear Mixed-effects Model Outcome of Mechanics Scores. 157
Figure 12 Performance of Feedback Groups in Terms of the Five Writing Components (Content, Organisation, Vocabulary Use, Language Use, Mechanics). 160
Figure 13 Performance of Teacher Feedback Group in Terms of the Five Writing Components. 158
Figure 14 Performance of Automated Feedback Group in Terms of the Five Writing Components (Text Content, Text Organisation, Vocabulary Use, Language Use, and Mechanics). 159
Figure 15 Performance of Hybrid Feedback Group in Terms of the Five Writing Components (Text Content, Text Organisation, Vocabulary Use, Language Use, and Mechanics). 160
Figure 16 Plot of the Linear Mixed-effects Model Total Writing Scores in the Two Writing Genres Tested (Expository, Persuasive). 169
Figure 17 A Proposed Plan for Integrating AWE into EFL Writing Instruction. 193

1. Introduction

1.1 Rationale and Background

Writing has its weigh in English language classrooms as one of the skills that English language learners should master, along with reading, listening and speaking. In a globalized world, proficiency in English writing has become a necessity in different segments of society. is a requirement to proceed in high level of education. Standardized English tests like TOEFL and IELTS are the key to be accepted in most of the universities. Appropriate writing skills are also a necessity in obtaining a good career. It is the case even for English native speakers as there is always a link between proficiency in writing and education.

Many studies have been conducted to examine a variety of methods related to writing as it is an important skill L2 learner must develop. One dominant phenomenon today is the integration of technology and computer software in English language classrooms. Thus, research is needed to validate the effectiveness of different types of software towards improving learners' skills and attitudes. For teachers of English L2 writing, the most significant question is how technology can be incorporated into teaching and how it could contribute to the final product of student writing. To that end, the current study will examine feedback generated by an automated writing evaluation system and its impact on the progress of students' writing. Specifically, it investigates how varying degrees of technology integration in the feedback process can be effectively incorporated into EFL classroom instruction.

Effective feedback, as many previous studies have confirmed (Ferris, 2003; Hyland & Hyland, 2006a, 2006b; Warschauer & Ware, 2006) is meant to enrich and improve learners' writing and should be comprehensive and compatible with their objectives and needs.

providing detailed and tailored feedback requires a well-trained and skillful instructor(Warschauer & Ware, 2006). Hence, instructors should be involved in a focused training to be able to provide effective feedback. However, providing sufficient feedback is a very time-consuming task for busy teachers (L. Rudner & Gagne, 2001). To obtain the best result teachers should produce individualized representative feedback for each and every student in a timely manner. On the other hand, teachers have other tasks that should be addressed during teaching, which may delay responding to students' writing. This delay in providing feedback may reduce its effectiveness and make students lose interest in reading and responding to the teacher's comments (Grimes & Warschauer, 2010). As a result, in order to avoid such obstacles, computers generated feedback presents itself as a possible solution.

1.2 Context of the Study

The current study was conducted at the English Language Institute (ELI) of the University of Jeddah, Saudi Arabia. It targeted foundation-year students enrolled in a pre-intermediate level English course during the first semester of the 2022/23 academic year. The participants were female EFL learners aged between 18 and 19.

The education system in the Kingdom of Saudi Arabia (KSA) is segregated by gender. Males and Females study separately in public schools during all their schooling years. There are some exceptions in private and international schools which apply a mixed-gender structure, though classroom settings remain segregated in accordance with religious and cultural requirements.

At university level, gender segregation continues, except in certain fields that require practical training or fieldwork, such as medicine and engineering. Despite this separation, the

quality of education remains unaffected, as all students follow the same curriculum, adhere to the same educational system, and take the same examinations.

English language instruction was initially introduced as a mandatory subject in public schools (Grades 1–12), beginning in year four (age 9). Private and international schools incorporate English into their curriculum as early as at a preschool stage. However, in the first semester of the 2021–2022 academic year, the Ministry of Education revised its policy to introduce English instruction from the first grade of primary school. This change underscores the growing importance of English as a foreign language in Saudi Arabia.

Higher education in Saudi Arabia is accessible to students who successfully complete Grade 12 examinations. The students then have a choice to enroll in one of the 29 registered state universities, 38 (established and licensed) private universities and colleges geographically distributed across the different regions of KSA (MOE, 2025). Admission to those universities and colleges is contingent upon meeting the required registration requirements. Additionally, students have the opportunity to apply for government-sponsored scholarships to study abroad. These overseas scholarship programs are available to both male and female citizens who meet the specified eligibility requirements.

University of Jeddah, the university at which the study was conducted, is one of the newest universities in Saudi Arabia. It was established in 2014 within the framework of the university's directives and the National Transformation Program which aligns with the promising vision of the Kingdom of Saudi Arabia 2030. The university offers a diversity in specialized programs that meet the needs of the job market and contribute to the preparation of future leaders.

University of Jeddah has a special English language programme that is organised and operated by the English language institute ELI at the university. Students in the foundation year at University of Jeddah are assigned to different levels of proficiency in English, according to their performance in high school (Grade 12) exit exams. Then they proceed to the next level, after passing the final exam at the current level. The English language courses delivered at the ELI were for beginner, pre-intermediate, and intermediate levels.

Intensive English course includes 18 hours per week over 12 instructional weeks focusing on developing the four skills of reading, writing, listening and speaking, as well as developing critical thinking skills and presentation skills. Grammar and vocabulary are taught in the context of authentic reading and listening texts, which use the target language in natural and appropriate linguistic context. The intended outcome is to develop the linguistic competencies needed to be able to communicate and interact in simple and familiar contexts.

Beginner course aims at helping learners to achieve an overall English language proficiency of beginner Basic User, defined as A1 level to reach A2 level on the Common European Framework of Reference (CEFR), developing students' ability to express themselves in simple, basic language and engage in an increasing range of social situations. Pre-intermediate course aims to help students at A2 'Waystage' level achieve an overall English language proficiency of low B1 'Threshold' level on the Common European Framework of Reference for Languages (CEFR). Intermediate course is designed to support learners at the mid-B1 "Threshold" level in progressing to the mid-B2 "Independent" level of English language proficiency, as defined by the Common European Framework of Reference for Languages (CEFR), thereby preparing them to successfully enter their chosen English-medium majors.

The English program curriculum is structured around clearly defined learning outcomes, and end-of-module assessments are designed to measure students' attainment of these goals. Instructional materials are aligned with CEFR level descriptors to ensure consistency with target proficiency levels. Each level is supported by a designated textbook (Life), developed in collaboration with National Geographic Press, intended to guide students through a year-long progression from beginner to intermediate proficiency.

Classrooms are equipped with modern technological tools, including computers, smart boards, overhead projector, continuous high-speed internet and digital resources, to support interactive language instruction. Instructors are encouraged to promote a communicative and engaging classroom environment to boost language learning.

1.3 The Use of Technology in the EFL Classroom and the Vision for Saudi Arabia

The Saudi education system actively encourages the integration of technology into classroom practices across all subjects, including English as a Foreign Language (EFL). In line with the Kingdom’s broader digital transformation agenda, the Ministry of Education, in collaboration with the Saudi Data and Artificial Intelligence Authority (SDAIA), has recently published a national guidebook on the educational use of generative artificial intelligence (AI). This document reflects the government’s commitment to supporting technology-enhanced learning and to ensuring that emerging tools are employed in ways that are both effective and ethically responsible.

The guidebook emphasizes the potential of generative AI to improve the quality of education, increase efficiency, and raise awareness among teachers, students, and parents regarding the responsible use of such technologies. It outlines both the benefits and risks of adopting AI in education, highlights the tools that can be incorporated into teaching and learning, and provides detailed instructions for implementation at different educational levels. These initiatives directly support the objectives of Vision 2030, which prioritizes digital innovation, the adoption of modern pedagogies, and the development of a knowledge-based society.

At the institutional level, universities across the Kingdom, including the University of Jeddah—the context of this study—actively promote the use of technology to enhance the learning process. Within the EFL classroom, teachers and students are encouraged to employ a range of digital tools and platforms that foster more interactive and effective learning experiences. For instance, both teachers and learners are provided access to multiple technological resources such as Blackboard, institutional email accounts, and curriculum-specific websites (e.g., *National Geographic Learning*: https://eltngl.com), with personal usernames and passwords to facilitate use. These resources support communication, access to authentic materials, and the integration of digital content into classroom practice.

In this context, the study participants—EFL teachers and students—are situated within an educational environment that actively supports and encourages the adoption of technology as a central component of teaching and learning.

1.4 Personal Motivation

My motivation to investigate automated feedback on writing in the Saudi Arabian EFL context is rooted in personal and professional experience. I hold an MA in Applied Linguistics and Teaching English to Speakers of Other Languages (TESOL) and have taught English in Saudi higher education for several years. This research is driven by a desire to contribute meaningfully to both my students—EFL learners—and my colleagues—English language teachers.

Throughout my teaching experience, I have recognized the critical importance of writing skills and the challenges students face in developing them. Writing is often one of the most difficult language skills for learners to master. In my classes, I have observed that when students are assigned writing tasks to complete at home, they frequently seek assistance from others or, in some cases, pay someone to write on their behalf. They then attempt to memorize

the text for use in exams. However, this approach is generally ineffective, as it becomes evident to teachers when a submitted text does not reflect the student's actual abilities. At the same time, a large number of student papers requiring individual feedback creates a substantial workload for teachers. Providing detailed, personalized feedback that supports student progress and improvement is time-consuming and often difficult to sustain at scale.

These challenges sparked my interest in exploring potential solutions, ultimately inspiring and shaping the focus of this research.

My initial exposure to the concept of automated feedback occurred during my MA studies in Applied Linguistics and TESOL at Newcastle University in the United Kingdom. I encountered the topic in a Computer-Assisted Language Learning (CALL) module led by Dr. Scott Windeatt. I was immediately drawn to the subject, particularly because of its connection to the integration of technology in language classrooms—an area that has long interested me.

I have a strong belief in the transformative role of technology in education. While I hold deep respect for the irreplaceable role of teachers, I also recognize that technological tools are increasingly shaping educational practices, much as they have influenced other aspects of daily life. Motivated by this interest, I began exploring the validity and reliability of automated scoring systems. Although I found that automated scoring still presents certain limitations, I also recognized its potential.

This exploration sparked a deeper professional purpose: I aspired to contribute meaningfully to improving the quality of writing instruction in higher education. Specifically, I wanted to actively address the ongoing challenges students encounter in developing strong academic writing skills. By exploring the potential of automated feedback, I sought to support

both learners and instructors in contributing to writing development in ways that are scalable, effective, and aligned with students' long-term academic and professional needs.

1.5 Purpose and Research Questions

The literature underscores the evolving role of automated writing evaluation (AWE) systems in writing instruction. Numerous studies suggest that AWE tools can offer significant benefits, particularly when used in combination with teacher feedback (Link et.al., 2020; Wang, 2019; Wilson and Czik, 2016). However, there are still important areas that require further research to paint a complete picture about the use of automated feedback in EFL writing classrooms.

Research by Wilson and Czik (2016) and Link et al. (2020) , for example, suggested that combining AWE tools with teacher feedback allows teachers to focus on higher-level feedback, while still benefiting from the time-saving features of automated systems. Similarly, Chandler (2003)and Dikli (2010) highlighted the limitations of AWE systems in providing nuanced and detailed feedback, particularly in content development and higher-order writing skills, where teacher feedback is often more effective. Similarly, Wang (2019) examined the combined use of teacher evaluation and intelligent computer-based essay assessment, finding that this hybrid approach outperformed traditional teacher-only feedback in supporting English writing instruction. Overall, these findings suggest the advantages of a hybrid approach, yet they also signal a continuing need to determine how best to balance the strengths of automated systems with those of teacher feedback.

Recent studies have also focused on comparing AWE feedback to teacher feedback in terms of assessing writing quality by each of them separately, rather than examining the integration of automated feedback as a blended or hybrid feedback model (Chandler, 2003; Dikli, 2010; Dikli and Bleyle, 2014). This oversight fails to address the potential advantages of

combining the two, which has been argued to be the most effective way to compensate for the limitations of each approach when used separately (see 3.10).

Moreover, studies investigating the implementation of hybrid feedback conditions (automated feedback + teacher feedback) have produced contradictory findings (Fan, 2023; Sari & Han, 2024). Fan (2023) examined 67 university EFL students, with one group receiving two rounds of Grammarly feedback alongside teacher feedback and a comparison group receiving teacher-only feedback. The findings revealed no significant differences between the two groups in terms of syntactic and lexical complexity, accuracy, or fluency. In contrast, Sari and Han (2024) investigated a combined feedback model in which students received automated feedback on sentence-level errors together with teacher feedback on content and organization. Their results showed that the hybrid condition not only enhanced students' writing performance but also improved their self-efficacy, self-regulation, and reduced writing anxiety. Moreover, qualitative data indicated that students held favorable perceptions of receiving both automated and teacher feedback.

This inconsistency creates a gap in understanding how the three feedback conditions—AWE-only, teacher-only, and hybrid feedback—differ in their impact on student writing performance. Addressing this gap is essential for determining the most effective feedback approach in EFL writing instruction.

Furthermore, most studies to date have focused on an overall writing improvement or a specific feedback type (e.g., high-level versus low-level feedback), without delving deeply into how AWE and teacher feedback might differently affect specific linguistic features, such as: text content, text organization, vocabulary use, language use, and mechanics. Weigle (2013) and Ware (2011) suggest that AWE systems may be more effective when used for mechanical aspects of writing, while teachers are better suited to address content and fluency issues.

Nevertheless, further research is needed to systematically compare how AWE and teacher feedback in isolation and combined may influence these distinct writing components over time.

Another area that remains underexplored is the role of genre and how it might impact the quality of feedback provided. Much of the existing literature examines general writing skills without accounting for how different writing genres (e.g., expository vs. persuasive) might interact with the type of feedback provided to affect students' writing outcome. Genre-specific writing requires different skills and focuses, and it is possible that AWE systems or teacher feedback may be more effective when used in some specific genres, but not in others. As such, there is a need for studies that assess the impact of feedback types across various genres to determine whether the integration of AWE into writing instruction should be tailored to specific writing tasks.

To address these gaps, the current study compares three feedback conditions: (1) automated feedback only, (2) teacher feedback only, and (3) hybrid (automated+teacher) feedback. By doing so, it aims to provide a more comprehensive understanding of how different

feedback conditions may impact on writing improvement. Moreover, the current study examines specific aspects of writing (text content, text organization, vocabulary use, language use, and mechanics) to assess which areas benefit most from the feedback provided. Lastly, the research explores how feedback effectiveness may differ across writing genres (expository and persuasive), thereby contributing to a more nuanced understanding of how AWE systems can be integrated into diverse writing instruction contexts. This will not only address current gaps in the literature but also provide practical implications for educators seeking to optimize the use of AWE tools in their classrooms.

The current investigation is led by the following research questions:

RQ1: To what extent do three different feedback conditions (teacher only, automated only, hybrid feedback) affect overall L2 writing production?

RQ2: To what extent do three different feedback conditions (teacher only, automated only, hybrid feedback) affect different components of L2 writing examined: text content, text organisation, vocabulary use, language use and mechanics?

RQ3: Do the effects of three different types of feedback on writing differ depending on the genre of writing?

1.6 Structure of the Thesis

This thesis consists of seven chapters. Chapter One presents the rationale for the study and provides background information on the use of feedback in writing. It also describes the context of the study, outlines the personal motivation behind the research, states the aim of the study, and introduces research questions that guide the investigation.

Chapter Two provides a review of the relevant literature related to the theoretical frameworks that underpin the study. It explores key concepts related to feedback in second

language (L2) writing, including various types of feedback, their pedagogical significance, and theoretical perspectives such as the interaction hypothesis, sociocultural theory, and the roles of mediation and the zone of proximal development (ZPD). The chapter also discusses additional language learning constructs, including input, output, noticing, and task complexity. Finally, it reviews the integration of technology in L2 writing instruction and its relevance to feedback practices.

Chapter Three introduces Automated Writing Evaluation (AWE) systems, discussing their definition, significance, and role in language learning. It provides an overview of various AWE tools used in language classrooms and delves into empirical studies that have investigated the use of AWE as a scoring tool. The chapter then reviews literature comparing the impact of AWE feedback on writing quality with that of teacher feedback. Finally, it examines studies that explore the integration of AWE feedback with teacher feedback.

Chapter Four outlines the research methodology employed in the study. It describes the research design, participant selection, data collection instruments, procedures, and ethical considerations. The chapter also details the data analysis methods, including the use of linear mixed-effects models to evaluate the impact of the different feedback conditions.

Chapter Five presents the quantitative results of the study, while Chapter Six provides a summary and a discussion of the findings in relation to the research questions and existing literature.

Finally, Chapter Seven offers concluding remarks and discusses the limitations of the study. It also highlights the theoretical, methodological, and pedagogical contributions of the research and provides suggestions for future research based on the study findings and limitations.

2. Literature Review: Theoretical Foundation

2.1 Overview

This chapter presents a review of the literature underscoring the theoretical foundation relevant to the proposed study. It begins with a general introduction to feedback on writing, discussing its importance as a tool for improving writing quality and fostering learner development. Following this, the chapter explores two prominent language learning approaches that underpin the use of feedback in L2 writing: the interaction hypothesis and the sociocultural theory. Additionally, it briefly examines the input, output, and noticing hypotheses, highlighting their relevance to the feedback process in L2 writing. The discussion further extends to the concepts of mediation in language learning and the zone of proximal development (ZPD), emphasizing their relevance to the feedback process. Finally, the chapter examines the growing use of technology in L2 writing instruction, specifically focusing on the human-computer interaction that informs the use of automated feedback as a learning tool in the current study. The chapter concludes with a summary of the key points discussed.

2.2 Feedback on Writing: An Introduction

Feedback on writing has been extensively investigated through research, particularly in the latter half of the 20th century (Hyland, 2019; Hyland & Hyland, 2006b). Early studies in the 1970s, spurred by the process writing movement, focused on the question of whether feedback should be provided at all, with researchers debating its effectiveness in enhancing student writing. Instructors typically viewed feedback as a critical part of writing instruction, providing guidance to help students not only improve specific pieces of writing but also develop as writers over time. Consequently, feedback became a central pedagogical tool in writing instruction, evolving in its form, purpose, and delivery.

Feedback on writing generally refers to the instructor's comprehensive comments as a reader, aimed at directing students to improve the overall quality of their writing (Hyland & Hyland, 2006a). These comments typically target multiple features of writing, including ideas, content, organization, coherence, lexical choices, argumentation, and mechanics. Rather than being restricted to grammatical correction, feedback on writing is holistic, addressing both higher-order concerns (such as the clarity of ideas or the strength of an argument) and lower-order concerns (such as spelling or punctuation). It can be delivered in a variety of formats—written, oral, or through computerized systems, as educational technology advances (Ferris, 2003).

In the teaching and learning context, Butler and Winne (1995)defined feedback as "information with which a learner can confirm, add to, overwrite, tune, or restructure information in memory, whether that information is domain knowledge, meta-cognitive knowledge, beliefs about self and tasks, or cognitive tactics and strategies" (p. 5740). Similarly, Hattie and Timperley (2007) stated that feedback is information provided to students about their performance or understanding after a learning experience. This feedback can be offered by a teacher, a peer, a book, or a parent, and is aimed at improving specific skills.

Several researchers (e.g., Bitchener, 2008; Evans et al., 2011) further postulate that feedback in the classroom environment is beneficial for learners. Bitchener (2008) found that students who received written corrective feedback demonstrated improved writing accuracy in the immediate post-test, and this level of performance was retained two months later when compared to a control group. Similarly, Evans et. al., (2011), indicated that students who received traditional process writing instruction experienced some declines in linguistic accuracy

while those who received dynamic WCF showed significant improvement in the linguistic accuracy of their L2 writing.

Given the broad scope of feedback in education, it is essential to clarify the distinction between feedback on writing and written corrective feedback (WCF), as the two are sometimes conflated. Feedback on writing is comprehensive and addresses various aspects of writing, such as content, organization, and argumentation. In contrast, WCF focuses specifically on grammatical accuracy and language correctness. Van Beuningen (2010) highlighted the role of WCF in second language acquisition (SLA), suggesting that error correction helps learners internalize linguistic rules and avoid repeated mistakes. From this perspective, WCF is not merely a form of remediation but a valuable tool for language development.

However, the effectiveness of WCF remains a contested issue. Truscott (2007), famously argued against the use of corrective feedback in writing, particularly in second language contexts, claiming that it may have detrimental effects. According to Truscott, an overemphasis on error correction can increase anxiety among learners, making them more apprehensive about making mistakes. This heightened anxiety, in turn, may inhibit learners' ability to engage meaningfully with the writing process, ultimately stifling creativity and reducing the overall quality of their work. Moreover, Truscott and Hsu (2008) stated that successful error reduction during the revision stage does not necessarily indicate actual language learning. However, Truscott acknowledged that revision based on both content and form plays a central role in effective writing. As a result, the question of whether WCF contributes positively or negatively to L2 writing improvement remains debated and is a controversial topic in contemporary research.

The present study focuses rather on feedback on writing (FoW), which is a broader in scope than written corrective feedback (WCF) and is concerned with how feedback contributes to the development of student writing as a whole. While WCF typically targets linguistic accuracy at the sentence or word level (e.g., grammar, vocabulary, and mechanics), FoW encompasses both discourse-level concerns (e.g., cohesion, organization, genre conventions, and idea development) and attention to language forms. In this sense, WCF can be considered a component of FoW, but FoW extends beyond error correction to address higher-order aspects of writing. As Yu (2021) notes, comprehensive feedback on writing can serve as an integral medium for increasing students' understanding of academic language and writing genres. However, it is worth noting that both WCF in a narrow sense and FoW stem from the same or closely related theoretical foundations. These foundations revolve around interaction in diverse contexts with emphasis on varied sources of learning, including cognitive, social, and sociocultural perspectives.

The research exploring the impact of comprehensive feedback on writing began gaining prominence in the 1970s, during the rise of the process writing movement (Hyland & Hyland, 2006b). At that time, the focus shifted from final products to the writing process itself. Feedback became a tool to guide student revisions throughout this process rather than merely offering evaluative comments after the writing was completed (Hyland & Hyland, 2006b). Feedback in this context encourages students to view writing as an iterative process, where drafts are revised and improved upon with the help of constructive comments from instructors or peers.

Feedback on writing can take various forms depending on instructional goals and the needs of

individual learners. For example, feedback can be formative, where the primary goal is to improve students' writing skills over time by offering constructive comments throughout the writing process. Formative feedback is typically more detailed and developmental, designed to guide students through the revision process. Hyland and Hyland (2006b) defined formative feedback as a practice that points forward to the learners' future writing and the development of their writing process. On the other hand, summative feedback is often given at the end of an assignment and tends to focus on evaluating the overall quality of the final product rather than guiding future revisions (Black & Wiliam, 1998). In fact, over the past decades, changes in writing pedagogy and research have affected feedback practices. For instance, summative feedback has gradually been replaced by formative feedback (Hyland & Hyland, 2006b) indicating its impact on the development of students' writing.

Although feedback is a crucial component of EFL writing classrooms worldwide, research remains inconclusive regarding the most effective practices for delivering it—whether through instructors, peers, automated systems, or a combination of these methods. Understanding the theoretical foundations that inform feedback practices is essential to addressing this uncertainty. Feedback strategies in EFL contexts have been heavily influenced by various theoretical frameworks such as the sociocultural theory, the interaction hypothesis, and the integration of technology in language classrooms. The following section will explore these theories and their implications for feedback on writing.

2.3 Importance of Interaction

Interaction is considered as a medium through which L2 learning can take place, rather than merely a form of practicing specific linguistic features (S. Gass, 2003). This perspective on

the role of interaction as a medium for language learning was emphasized in Long's (1996) revised version of the interaction hypothesis. According to this, interaction may serve as an initial step in language learning (Gass, 1997). Long (1996), demonstrated that interaction with a native speaker or a more competent interlocutor triggers linguistic adjustments and facilitates acquisition by connecting input and output in a productive way. Ellis (1999), also argued that

interactionally modified input, which occurs as needed during goal-oriented conversation, is more effective for learning a target word or structure than pre-modified input, which is adapted to be target-like before any learner error occurs. This emphasizes that interactionally modified input promotes more effective acquisition in the context of language learning.

The interaction hypothesis is considered one of the prominent frameworks in the study of language acquisition, associated with various language learning hypotheses rather than being a standalone theory itself (Mackey, 2012). It is usually integrated with other approaches to language learning such as the input hypothesis (Krashen, 1982, 1985), the output hypothesis (Swain, 1985, 1995), the noticing hypothesis (Schmidt 1990, 2001). Long (1996) suggested that second language learning is facilitated through interactional processes as interaction plays a crucial role in linking "input, internal learner capacities, particularly selective attention, and output in productive ways" (Long, 1996, p. 451).

Additionally, the interaction hypothesis emphasizes the importance of the social context and other cognitive processes such as the working memory and aptitude in shaping learning outcomes (Mackey, 2012). Pica (1996) further elaborated that the interaction hypothesis is integrally linked to several language acquisition theories, solidifying its role as a comprehensive and multifaceted model that supports both cognitive and sociocultural perspectives on language learning. Interaction, thus, can be considered as a facilitator of many of the processes involved in learning (Mackey, 2012).

Even though the interaction hypothesis was formulated with the aim to promote L2 development in speaking skill, this theoretical approach is equally applicable to the writing medium. Especially, when considering the role of feedback, the interaction with an interlocutor

providing feedback can easily find its parallel in learner interacting with the feedback on writing given in a written form.

Furthermore, just as interaction in spoken language fosters improvement through immediate feedback, written feedback plays a crucial role in developing writing skills. Mackey (2012) demonstrated that engagement with written feedback, as oral feedback, creates an implicit dialogue in which learners interact with the feedback provided through revisions or questions. The process of receiving and responding to feedback supports learners in refining their writing, much like oral interaction helps them adjust their spoken language. The iterative nature of engaging with written feedback is often part of a larger cycle of drafting, revising, and improving written output (Mackey, 2012). This supports the critical role of feedback in both oral and written language development.

The interaction hypothesis, thus, can be applied to various types of interactions. For example, in Wilson and Czik's (2016) study on the effect of automated feedback on teacher feedback and writing quality, multiple forms of interaction were evident, including learner-AWE system, teacher-AWE system, teacher-learner, and learner-learner interactions. Students engaged with both the AWE system and teacher feedback by receiving immediate feedback on their writing, revising their work accordingly, and resubmitting it for further evaluation. Teachers also interacted with both the system and students, clarifying certain aspects of the AWE feedback and providing their own comments on learners' writing. Additionally, interaction occurs between learners, allowing them to evaluate and provide feedback on their peers' work. This highlights that interaction can take place in various modes and environments, underscoring the importance of examining its effects beyond oral interactions.

The relevance of the interaction hypothesis to the current study is evident in how learners in both the AWE and teacher feedback groups engage with written feedback. When AWE is delivered during synchronous computer-mediated communication (SCMC), it allows immediate clarification and negotiation of meaning, which reflects the core principles of the interaction hypothesis (Long, 1996). By contrast, teacher FoW is more often provided asynchronously (e.g., teacher comments on drafts), where opportunities for real-time negotiation are limited. Nevertheless, the interaction hypothesis remains relevant to the present study, as learners in both the AWE and teacher feedback groups are prompted to interact with the feedback they receive—whether through dialogue with a teacher or engagement with a system—to revise their input and produce improved output. These asymmetrical interactions, between learner and teacher or learner and computer, are vital for making input comprehensible and providing learners with feedback on their writing. Such feedback, especially when it focuses on form, meaning, or both, is considered essential for language learning. By addressing specific areas for improvement, these interactions may facilitate the development of L2 learners' written text quality. Thus, whether through synchronous exchanges or delayed engagement with feedback, both types of interaction—learner-teacher and learner-computer—offer valuable opportunities for learners to develop their writing skills.

This understanding of feedback as an interactive process underscores the role of different modes of feedback in supporting L2 writing development. Interaction in the form of written feedback has been shown to have positive implications for language learning, both in providing comprehensible input to language learners and in facilitating the production of improved output, i.e., modified output, in the target language.

However, to fully understand the effectiveness of feedback in L2 writing development, it is essential to compare different modes of feedback and their impact on EFL learners' writing improvement. While previous research has examined the role of oral and written feedback, limited attention has been given to how various types of written feedback—such as teacher feedback, and automated feedback—differ in their influence on writing quality and learner's engagement. By investigating these differences, this study aims to provide deeper insights into the most effective feedback mechanisms for improving EFL learners' writing skills, contributing to a more nuanced understanding of how interaction facilitates language learning.

2.4 Input, Output and Noticing

Linked to the interaction hypothesis, Krashen's (1978) input hypothesis suggests that language learning occurs when learners are provided with comprehensible input. However, input is not assumed to be beneficial on its own (Mackey 2012).Long (1981) contended that modified input can be beneficial to language learning only when it is combined with either modified or unmodified interaction. Without proper interaction with the provided input, there is a risk that the intended language learning may not occur.

A further interactional process that can result from feedback is known as modified output (Swain 1985, 1995). Swain encouraged researchers to view learners' output as an integral part of language acquisition. She argued that forcing language learners to produce language is the key that may shift their focus from merely understanding the meaning of the input to prioritizing syntactic accuracy (Swain 1985). According to Mackey (2012), output or merely producing the language may not help language learners to acquire the language, it is rather considered as a form of practicing. Language production then should be directed in a way in order to help acquisition. There should be a form of feedback to help learners in producing comprehensible output. Comprehensible output refers to the need for a learner to be "pushed toward the delivery

of a message that is not only conveyed, but that is conveyed precisely, coherently, and appropriately" (Swain, 1985, p. 249).

As much as the importance of feedback in oral interaction, feedback on writing can contribute to the improved production of language in the written medium. Empirical research within the interaction hypothesis framework has also demonstrated that interaction, which pushes learners to extend their linguistic resources and modify their output in response to feedback, can facilitate the development of certain linguistic forms (Long, 1996) or, at the very least, increase their awareness of specific linguistic structures (Ellis, 1999).

The other crucial approach to SLA that is facilitated by interaction through feedback is the noticing hypothesis (Schmidt, 1990; 2001). Noticing acts as the cognitive mechanism that allows learners to recognize and attend to linguistic forms, which in turn facilitates language development. Schmidt (1990) argued that noticing —explicit awareness— is essential to language acquisition. According to the noticing hypothesis, language acquisition occurs when conscious attention enables learners to process input and convert it into intake. In this regard, "intake" represents the portion of input that is noticed, attended to, internalized and made available to contribute to language development (Ellis, 1994). This is particularly relevant in feedback driven interaction, where feedback helps learners notice gaps between their interlanguage and the target language.

Feedback provides comments that explicitly direct learners' attention on the gap between their oral or written production and the more proficient speaker's production (Schmidt 2001). Ellis (1995) refers to this notion as cognitive comparisons, where learners must notice and compare the input with their own output. However, Schmidt (2001) stated that attentional resources in the human cognitive system are limited. When individuals are presented with multiple stimuli simultaneously, the brain may struggle to process all of them effectively due to

its limited processing capacity ((Skehan, 1998; Al-Hejin, 2004).

2.5 Task Complexity

There has been a longstanding debate between Skehan's limited capacity hypothesis (LCH) and Robinson's cognition hypothesis regarding the development of complexity as a component of language proficiency. Skehan's LCH proposes that learners have limited attentional resources that must be distributed among different dimensions of language performance—such as accuracy, fluency, and complexity—making it difficult to improve all aspects simultaneously during complex tasks.

As explained by the limited capacity hypothesis (Skehan, 1998), in the context of second language learning, this limitation can lead to cognitive overload, particularly when learners are required to divide their attention between real-time language production, comprehension, and verbally delivered feedback. Excessive demands on attentional resources—such as processing feedback while actively constructing sentences—can overwhelm learners. This suggests that effective feedback strategies should consider cognitive load, ensuring that learners are given opportunities to process and incorporate corrections without excessive strain on their attentional capacities. Written feedback provides learners with sufficient time to interact with the feedback, compare their output with the input, and make necessary revisions. This extended processing time is likely to help learners' ability to notice gaps in their interlanguage (Polio et al., 1998; Sheen, 2011).

While Skehan applied his framework primarily to oral production (drawing on Levelt's model), the underlying principle of attentional constraints is relevant to writing as well. Writing similarly requires the simultaneous management of multiple cognitive processes, including planning, translating ideas into text, and revising (Kellogg, 1996; Révész et al., 2017).

Kellogg's (1996) writing model provides a more targeted framework for understanding

how learners manage these processes. It outlines the cognitive demands of writing tasks, including planning, translating ideas into text, and reviewing/revising, and highlights the limitations of attentional resources when learners process feedback concurrently with text production. Incorporating Kellogg's model allows a clearer connection between task complexity and the cognitive processing of feedback in writing, complementing Skehan's emphasis on attentional limitations.

According to Skehan and Foster (2012), an L2 learners may not be able to provide equal attention to fluency, complexity, and accuracy simultaneously due to limitations in attentional resources. As a result, learners often prioritize completing the task over focusing on language accuracy or development. LCH suggests that because learners have a restricted cognitive capacity for processing information, they are thought to allocate their attention selectively. This may involve either concentrating on the content of the task or focusing on linguistic form to support language learning (Tavakoli, 2009).

In contrast, Robinson (2001) rejected the idea of limited attentional capacity and argued that increased cognitive complexity can lead to the production of more accurate and more complex language output. Robinson's cognition hypothesis (2001, 2007) posits that cognitively demanding tasks, particularly those that increase demands along resource-directing dimensions, promote greater attentional allocation, deeper memory encoding of input, more differentiated language use, and heightened noticing of mismatches between learner output and target forms. As Robinson (2012) explains, "the greater cognitive demands of complex tasks along resource-directing dimensions will lead to greater attentional allocation to, and rehearsal of, input in memory; greater functional differentiation of language use; and also more extensive noticing of mismatches between learner output and target input" (p. 317). Robinson (2007) further contended that attentional resources are not inherently limited; rather, learners can access

multiple, non-competing attentional systems simultaneously. From this perspective, cognitively complex tasks do not overload the learner but instead facilitate the production of more complex language, often reflected in greater lexical diversity and grammatical accuracy.

Robinson (2001, 2012) also distinguished between task complexity, task difficulty, and task conditions. Task complexity refers to characteristics of the task itself that can be manipulated to increase or reduce cognitive demands during performance. Task difficulty, by contrast, is influenced by individual learner variables such as motivation, confidence, or aptitude. Task conditions involve the contextual and interactive elements of a task, including participant roles and interaction types. Among these dimensions, task complexity is particularly relevant to the cognitive demands of writing and is therefore central to this study. While much of Robinson's work focuses on spoken production, the underlying principles extend to writing, where cognitive load similarly affects the complexity and quality of learner output (Révész et al., 2017).

The current study draws on the cognition hypothesis and the concept of task complexity by comparing learner performance across two genres—expository and persuasive. Persuasive

writing is considered more cognitively demanding than expository writing, as it requires critical thinking, synthesis of multiple perspectives, and the construction of logical arguments. The study aims to explore which type of feedback—teacher, automated, or hybrid—is most effective for supporting learners across tasks of varying complexity. Specifically, it investigates whether feedback influences writing performance differently depending on the cognitive demands of the genre, and whether more complex tasks benefit more from certain types of feedback.

In addition to cognitive perspectives, the significance of feedback is also supported by **s**ociocultural **t**heory, which emphasizes the crucial role of social interaction in the learning process. To better understand how feedback functions within this framework, the following section discusses key principles of **s**ociocultural **t**heory and their relevance to language learning.

2.6 Sociocultural Theory

Sociocultural theory (SCT) also examines the role of interaction but within a broad cultural and social context (Warschauer, 1997), p.471). It illuminates the role of social interaction in creating an environment for language learning. SCT, originally developed by Vygotsky and Cole (1978), provide a theoretical framework that guides the current study besides the interaction hypothesis discussed in the previous section.

Vygotsky's work has influenced areas such as psychology and education. He emphasized the relationship between social interaction and cognitive development. Over the past decade, the concepts of learning and development have sparked mixed views among scholars. However, Vygotsky acknowledged that learning itself is not equivalent to development. He emphasized that a properly organized learning atmosphere can foster mental development and initiate a range of developmental processes that would be unattainable without such learning opportunities (Vygotsky, 1978). Furthermore, he argued that learning begins through social interactions, which

subsequently drive cognitive development. SCT, therefore, posits that knowledge is co-constructed in social contexts rather than being developed solely within an individual's mind.

According to SCT, learning occurs within a context shaped by interactions with others, particularly with more knowledgeable individuals, such as teachers or advanced peers ((Bornstein & Bruner, 2014; J. P. Lantolf & Thorne, 2006). These interactions foster higher-level cognitive abilities, including logical thinking, memory, and problem-solving, which are deeply rooted in the learner's social and cultural environment. In contrast to the cognitive approach, which views knowledge development as internal process that is constructed within individuals' minds, SCT emphasizes that all cognitive development, including language learning, is internalized within social settings and mediated by cultural artifacts ((Lantolf & Pavlenko, 1995; Swain & Watanabe, 2013). Lantolf and Thorne (2006) further highlighted that cognitive growth occurs in historically and socially situated contexts, such as family life, peer interactions, or educational institutions. The following section explores key concepts of sociocultural theory (SCT) that are relevant to understanding how feedback may contribute to L2 language learning.

2.6.1 Mediation in Language Learning

A key component of SCT is the concept of mediation. Vygotsky (1987) argued that human mental activity relies on tools and artifacts to mediate and regulate knowledge. Social relationships and artifacts determine the extent to which an individual's mental processes are developed ((Lantolf, 2000). Artifacts can range from simple physical objects, such as pen and paper, to more complex tools like computers. Through mediated interactions, humans utilize culturally developed physical and symbolic tools and artifacts**,** such as music, language, numbers, or art, to mediate their relationships with the surrounding environment (Vygotsky, 1987). Those mediated interactions not only facilitate human mental development but also

transform it. To illustrate the concept of mediation, Lantolf and Thorne (2006) used the example of digging a hole, highlighting how modern humans rely on tools like a shovel, which offer greater efficiency compared to using their bare hands. Figure 1 illustrates how mediation in SCT links the individual to their surrounding environment. Interaction with mediational tools in this context not only facilitates but also transforms individual thinking.

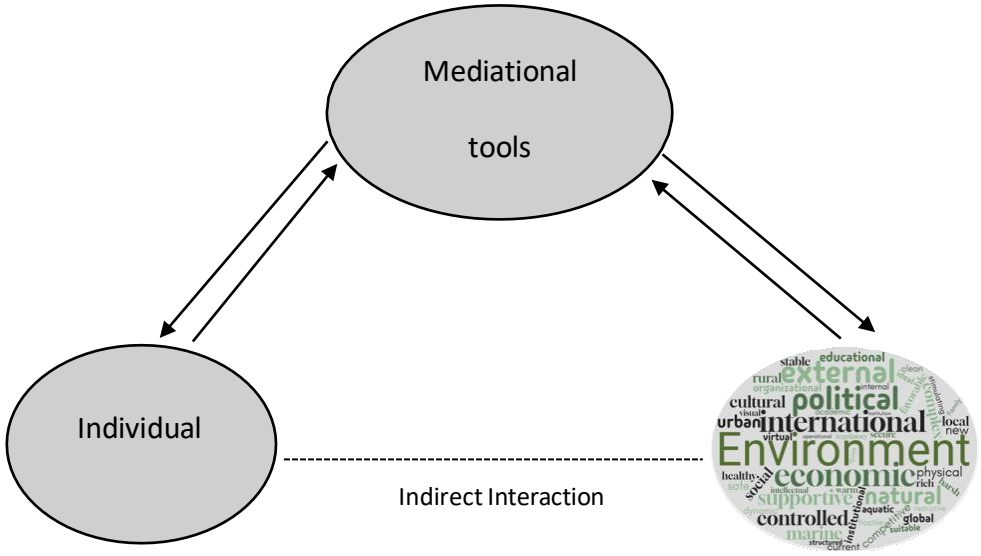


Figure 1 Mediation in Sociocultural Theory.

Regarding second language learning context, Lantolf (2000) further emphasized that second language learning is inherently a mediated process, drawing from the foundational principles of sociocultural theory. He identified three distinct categories of mediation involved in language learning: social mediation, self-mediation, and mediation through artifacts. Social mediation refers to the interactions between learners and more knowledgeable others, such as teachers, peers, or experts, who provide guidance and support in the learning process. This collaborative exchange facilitates the co-construction of knowledge. Self-mediation, on the other hand, involves the learner's ability to regulate their own cognitive processes, such as through private speech, self-reflection, or the strategic use of language to solve problems. Finally, mediation through artifacts highlights the role of cultural tools, both physical and symbolic, in shaping the learning experience. Examples of such artifacts include textbooks, technology, and even linguistic symbols, which learners use to process, internalize, and produce language. Together, these three forms of mediation underscore the dynamic and interactive nature of second language acquisition, demonstrating how external tools and social contexts are integral to the internalization of linguistic knowledge.

Mediation in SCT and feedback on writing are closely interconnected. Feedback on writing serves as a mediational tool that helps learners refine their writing skills. Additionally, feedback—whether provided by teachers, peers, or through self-reflection—acts as a mediational means that bridges the gap between a learner’s current level of performance and their potential for improvement. Social mediation, for example, occurs when instructors or peers guide learners by providing constructive comments on grammar, structure, or ideas. Self-mediation, on the other hand, is evident when learners use feedback to regulate their own writing processes. They may reflect on prior comments, adjust their strategies, and independently apply corrections to

future drafts. Lastly, mediation through artifacts can be seen when learners engage with written tools, such as rubrics, style guides, or automated feedback systems, which provide structured input to guide revisions.

By connecting SCT's mediation to feedback on writing, it becomes clear that feedback is not just corrective but also transformative. It enables learners to internalize new knowledge, develop higher-order cognitive abilities, and produce more sophisticated written work. Thus, feedback exemplifies the principles of SCT by serving as both a tool and a process that facilitates the co-construction and internalization of writing skills.

2.1.1 Zone of Proximal Development

Vygotsky introduced the concept of the zone of proximal development (ZPD) as a framework for understanding how social interaction can facilitate cognitive development. According to Vygotsky (1978), the ZPD is defined as "the distance between the actual development level, as determined by independent problem solving, and the level of potential development, as determined through problem solving under adult guidance or in collaboration with more capable peers." (Vygotsky, 1978, p.86). Lantolf (2000), further illustrated that the ZPD represents the difference between what individuals can achieve independently versus what they can accomplish with support from others or cultural artifacts. He also stated that ZPD is a collaborative construction of opportunities for individuals to enhance their mental abilities. In the ZPD, children or learners engage in activities beyond their current individual capabilities with the support and guidance of someone more experienced. This collaborative process not only facilitates learning but also fosters cognitive growth by bridging the gap between what learners can do independently and what they can achieve with assistance. Therefore, this suggests that

what a child can accomplish with support today, they will be capable of doing independently tomorrow (Ohta, 2000).

In developmental research, the ZPD has emerged as a powerful concept for diagnosing educational problems (Vygotsky, 1978). Lantolf and Poehner (2008) argued that Vygotsky's theory of development is relevant to all aspects of mental development, not only early childhood. In the context of second language acquisition (SLA), for instance, the concept of the ZPD has been successfully applied to stimulate language learning (Ohta, 2000; 2001). (Ohta, 2001) for example, has adopted the notion of the ZPD to explain how teacher or peer assistance is linked to language development. Notably, she found that in peer collaboration, the students were exchanging roles of novice and expert, yet they succeeded in improving the accuracy of language they produced.

Aljaafreh and Lantolf (1994) also supported the application of ZPD in language learning. However, they argued that specific guidelines should be followed to achieve the ZPD in a classroom setting. They asserted that the assistance provided by the expert to the novice should be gradual, offered only when necessary, and withdrawn once the novice can independently perform the task (Aljaafreh & Lantof, 1994). Furthermore, in interactions between experts and novices in teaching situations, novices do not simply copy the expertise of the experts; rather, they transform and appropriate the knowledge offered (Lantof, 2000). Other factors influencing the ZPD include the expertise of the helper, the nature of the task, the goals of the participants, and the developmental levels of the learners (Ohta, 2000).

It has been argued that the notion of ZPD can be considered as a form of interaction. Aljaafreh and Lantolf (1994) defined development in language acquisition context as "the study

of how mediational means are appropriated by the individual as a result of dialogic interaction with other individuals" (Aljaafreh & Lantolf, 1994, p. 467). Research based on SCT addresses corrective feedback as a collaborative process where the dynamics of the interaction itself shapes the nature of feedback and inform its usefulness to the learner (or learners in the case of more symmetrical peer-interaction).

The concept of the zone of proximal development (ZPD) is fundamentally tied to dialogic interaction. Aljaafreh and Lantolf (1994) described development in language learning as "the study of how mediational means are appropriated by the individual as a result of dialogic interaction with other individuals" (Aljaafreh & Lantolf, 1994, p. 467). This underscores the idea that learning is a collaborative process, where shared dialogue plays a pivotal role in fostering understanding. Research based on SCT views feedback as a cooperative process, with the quality and impact of feedback being influenced by the nature of the interaction itself (Lantolf & Thorne, 2006). This is particularly true in peer interactions, where the balance of knowledge is more symmetrical. However, the effectiveness of feedback depends on how closely it aligns with the learner's developmental stage within their ZPD. As noted by Aljaafreh and Lantolf (1994) feedback becomes relevant for learning when it addresses specific aspects of the target language situated within the learner's ZPD.

Vygotsky argued that the relationship between individuals and their environment is mediated in three primary ways: through interaction with more knowledgeable individuals, through the use of physical tools, and through symbolic artifacts such as language Poehner & Lantolf, 2013). Aljaafreh and Lantolf (1994) further proposed that interaction within ZPD should be graduated, contingent, and take the form of dialogue to effectively mediate learning within the

ZPD. They argued that dialogic exchanges are essential to uncovering and addressing the learner's developmental needs as it is challenging to identify the boundaries of the learner's ZPD and to provide appropriate support without meaningful dialogue. This focus on interaction highlights SCT's emphasis on the social and collaborative nature of learning.

In this context, the effectiveness of feedback relies on its timing and relevance to the learner's ZPD, which represents the distance between what learners can do independently and what they can achieve with guidance. To foster development, feedback must be closely aligned with the learner's specific needs in relation to the properties of the target language (L2) (Aljaafreh & Lantolf, 1994). From this sociocultural perspective, feedback on writing can be viewed as a form of mediation that operates within the learners' ZPD, offering scaffolding that supports writing development. By providing individualized, context-aware feedback, L2 learners can efficiently bridge the gap between their current abilities and their potential proficiency in the target language.

The following section addresses the use of technology in L2 writing instruction, specifically focusing on the human-computer interaction that informs the use of automated feedback as a learning tool in the current study.

2.2 The Role of Technology in Second Language Writing

Feedback on writing has been significantly transformed by the emergence of technology in language classrooms. With continuous advancements, technology has become indispensable for language learners worldwide, reshaping how feedback is provided and processed. Teachers and institutions face the ongoing challenge of integrating technology effectively, not only as a source of content but also as a means of facilitating authentic language learning experiences (Chapelle

& Sauro, 2017). According to Bax (2003), technology has reached a state of normality, seamlessly embedding itself into everyday educational practices. This integration has revolutionized second language learning particularly with the use of the internet.

In L2 writing, for example, teachers and learners now have access to vast online resources, real-time automated feedback that enables multiple writing revisions, and opportunities for independent self-development. With the use of technology, learners can refine their writing skills both inside and outside the classroom. Engagement with teacher and peer feedback become also possible through digital platforms, which facilitate both asynchronous and synchronous feedback exchanges**,** helping L2 learners improve their writing.

Various studies have examined teachers use of digital platforms to deliver feedback on learners' written assignments (e.g., Tuzi, 2004). Some of the platforms that frequently used by teachers were Google Docs, Microsoft Word Processor, automated writing evaluation tools. Hyland (2019) stated that teachers apply many forms to respond to students written work. For example, teachers' response could be in a form of commentary, rubric, minimal marking or electronic feedback. Tuzi (2004) defined electronic feedback (e-feedback) as "feedback in digital, written form and transmitted via the web" (Tuzi, 2004, p. 217). Delante (2017) listed some terminologies used in literature to refer to teacher technology-mediated feedback. Some of those terminologies are electronic feedback, teacher electronic feedback, Internet-mediated feedback and computer-mediated human feedback (Delante, 2017). Those multiple terminologies underscore the impact of technology on feedback transferring the concept of oral/written response into the electronic arena (Tuzi, 2004).

Technology serving as a mediating tool may fosters interaction and collaboration in writing classrooms context.Lantolf (2000b,a) argued that technology can be linked to sociocultural

theory as a mediational means through which language learners interact with native speakers, language instructors, or more advanced peers (Lantolf, 2000b). Loncar et al. (2023) also affirmed that teacher as well as peer feedback can be mediated through the use of technology. Within this framework**,** use of technology provides new opportunities for interaction, redefining traditional discourse structures.

Warschauer (1997; 1999) also applied sociocultural theory to guide his research on computer mediated communication **(**CMC) environments. He used a conceptual framework that starts with SLA theories of input (Krashen 1982) and output (Swain 1995) and leads to sociocultural learning theory (Vygotsky 1987) demonstrating how human interaction with technological tools shapes the learning process. A key finding from his work highlights the reflexive relationship between humans and technological artifacts, emphasizing that learners do not simply use technology but interact with it in a way that transforms their learning experience. CMC, hence, fosters a unique form of discourse that differs from face-to-face interactions, offering a more flexible and inclusive communication environment.

One of the drawbacks of written feedback, whether provided by teacher or peers, is being relatively slow compared to oral feedback. However, with the revolution of CMC, Warschauer (1997) argued that the computer-mediated features of writing online have unleashed the interactive power of text-based communication. He added computer-mediated text-based interaction is easily transmitted, stored, archived, reevaluated, edited, and rewritten (Warschauer, 1997, p. 472). Moreover, this kind of written feedback can provide a permanent record of teacher-student / student-students interactions that can be revisited and used to promote learning (Cummins & Sayers, 1995). These characteristics position technology as a critical tool for

elevating the process of written feedback, reinforcing the importance of exploring how different digital platforms shape written communication and improve writing skills.

The integration of technology in language learning, hence, supports Long's interaction hypothesis, as it creates opportunities for meaningful interaction through digital platforms. Chapelle (2017) stated that utilizing the internet, learning platforms, and social media facilitates interaction between language learners and their peers or teachers. She further referred to this type of interaction as Human-Human Interaction (HHI), where learners use computers as a vehicle for interacting directly with other language learners or teachers (Shannon & Chapelle, 2017). For example, language learners can engage in real-time chat programs to complete language tasks, reinforcing the role of technology-enhanced interaction in second language acquisition.

Warschauer (1997) compared face-to-face and online interaction and summarized key features of online communication as text-based and computer-mediated, time- and place-independent, and accessible across long distances. The asynchronous nature of online, text-based interaction is thought to be better suited for practicing complex writing and problem-solving tasks than synchronous discussions in a classroom setting.

Additionally, online interaction allows learners to engage in communication regardless of time and location and is accessible across distances, enabling them to benefit from technology both inside and outside the classroom. It also facilitates interaction with individuals worldwide, broadening exposure to diverse language use and cultural perspectives. These features make online interaction a potentially valuable tool for language learning. However, it is important to explore the impact of using technology in real language learning classrooms to validate its effectiveness (Warschauer, 1997).

In addition to Human-Human Interaction (HHI), another type of interaction that has emerged due to continuous technological advancements is Human-Computer Interaction (HCI) (Shannon & Chapelle, 2017). HCI occurs when learners engage independently with web-based software or multimedia for learning purposes. Technologies that support HCI can be categorized into three main types: Web 2.0 applications, automated writing evaluation systems, and corpus-based tools (Li et al., 2017). Each of these technologies allows learners to interact autonomously with digital tools promoting their learning experience.

Web 2.0 applications are one of the earliest technological advancements researched in the context of language learning. Examples of Web 2.0 applications include social media sites, Google Docs, and Microsoft Word processors. Social media platforms, such as Facebook, Twitter, Instagram, LinkedIn, and discussion forums, provide language learners with authentic content and real-world discourse. Collaborative tools like Google Docs also facilitate real-time collaboration, peer review, and interactive writing exercises, making them particularly useful for developing writing skills. Microsoft Word also encourages writing practice. It is one of the most widely used word processing tools that supports writing development. It provides some features like built-in spelling and grammar checks, which help learners notice errors and make corrections independently.

Automated writing evaluation (AWE) systems, such as Criterion, Turnitin, and Writing Pal, are also considered examples of Human-Computer Interaction (HCI). These platforms provide immediate feedback on grammar, style, and plagiarism, allowing learners to independently refine their writing. By using AWE platforms, learners engage with written text autonomously, revising and editing their work, which enhances both autonomy and language learning.

Another example of HCI in language learning is the use of corpus-based tools. These tools enable students to analyze large databases of authentic language use, helping them understand vocabulary usage, collocations, and language patterns in context. Additionally, corpus-based tools exemplify data-driven learning (DDL), a learning approach in which learners explore large datasets of authentic texts to identify and notice target linguistic forms, structures, or vocabulary patterns. This exploratory, evidence-based approach encourages inductive learning, fostering greater linguistic awareness and self-directed learning.

In Human-Computer Interaction (HCI), when using one of the previously mentioned tools, learners can engage independently with texts, fostering greater autonomy in editing, revising, and refining their writing. However, when using automated writing evaluation (AWE) tools, HCI and Human-Human Interaction (HHI) intersect. AWE tools provide timely feedback, allowing learners to engage with content autonomously (HCI). At the same time, they facilitate interaction with other users or instructors (HHI) by enabling peer and teacher feedback. This integration of automated and human-mediated feedback supports both independent learning and collaborative interaction, is thought to enrich the overall language learning experience and, more specifically, advancing writing skills.

2.3 Summary

Various language learning theories support the role of feedback in L2 learning, asserting that it has a positive impact on language development. Within the Interactionist approach, interactional feedback is recognized as a key mechanism for ameliorating input comprehensibility and facilitating the production of modified output.

From a sociocultural perspective, feedback—particularly from teachers and peers—is considered an essential mediational tool that supports language learning within the zone of

proximal development (ZPD). In this framework, more knowledgeable others (e.g., teachers or advanced peers) provide scaffolded feedback, enabling learners to progress beyond their current level of linguistic competence.

The integration of technology further supports the effectiveness of feedback in L2 teaching and learning. With advancements in automated feedback systems, learners can receive instant, data-driven corrections through technological tools, supplementing or replacing traditional teacher-provided feedback. Moreover, technology may facilitate learner interaction with teachers and peers in both asynchronous and synchronous environments and subsequently expand opportunities for engagement with feedback.

Despite the growing body of research on feedback on writing, the most effective approach to feedback implementation remains an area of ongoing investigation. There are still gaps in understanding how different written feedback modalities contribute to L2 writing development. Thus, the present study seeks to empirically examine and compare the impact of teacher-written feedback, automated feedback, and a hybrid feedback condition (combining teacher and automated feedback) on EFL learners' writing skills. By exploring these different feedback modalities, this research aims to contribute to the ongoing discussion on effective feedback practices in L2 writing instruction and provide insights into the pedagogical implications of technology-enhanced feedback.

The following chapter is intended to present representative empirical research relevant to the impact of using automated writing evaluation systems on writing quality.

3. Empirical Research on the Impact of AWE on Writing Quality

3.1 Overview

This chapter presents a review of empirical research focused on the implementation of automated feedback in L2 writing classrooms, particularly its impact on writing accuracy and quality, and how it compares to traditional teacher-written feedback. It begins by reviewing the definition and significance of automated writing evaluation (AWE) systems, emphasizing their role in supporting L2 writing development. Following this, the chapter discusses the development of AWE tools and provides examples of their use in L2 writing classrooms, highlighting how these tools have evolved to assist language learners. The effectiveness of AWE as both a feedback mechanism and a scoring system is then examined, exploring its accuracy and reliability in assessing writing. Next, empirical research comparing AWE feedback with teacher-written feedback is explored, offering insights into the strengths and weaknesses of both approaches. The chapter also addresses research relevant to the current investigation, focusing on the impact of combining AWE and teacher-written feedback on student writing. Additionally, it sheds light on the challenges associated with AWE implementation. A brief conclusion is provided to summarize the key findings from the reviewed studies and the gaps in the literature that the current study aims to address.

3.2 Automated Writing Evaluation Definition and Significance

A number of computer applications have been developed and investigated for assessing learners' writing products. These applications can provide both scores and formative feedback to improve writing quality. In the literature, these applications are primarily known as automated essay scoring (AES) or automated writing evaluation (AWE) systems. AES refers to the scoring engine used to generate scores for students' written essays. E-rater, for example, is one of the

scoring systems that is used in high- stakes tests like TOFEL, GRE, and GEMAT (Grimes & Warschauer, 2010).

The term automated writing evaluation (AWE) generally refers to the scoring and assessment of written texts by providing both summative and formative feedback. The Project Essay Grader (PEG) program, developed by Page in 1966, was the first attempt to provide feedback on written essays. Since then, several AWE programs have been created to offer human-like evaluations of students' writing, including My Access, Criterion, Essay Critique, iWrite, and others.

Shermis et al. (2013), defined automated writing evaluation (AWE) as "the process of evaluating and scoring written prose via computer programs." Azmi et al. (2019) also used the term "evaluation" to refer to computer programs that generate feedback on students' writing. He explained that these programs can provide commentary that contributes to improving students' writing. AWE tools not only score a given piece of writing but also offer individualized feedback on vocabulary, grammar, style, coherence, and discourse (Rudner & Gagne, 2001). The term "evaluation" encompasses a broader process that includes scoring, assessing written texts, and providing relevant feedback that writers can use to improve the quality of their work. Due to these features, the use of AWE applications by ESL students as formative tools in writing classrooms is increasing (Stevenson, 2016).

3.3 AWE Tools: Functions and Feedback Mechanisms

AWE tools comprise a variety of computerized methods to assign scores to the written texts. Except for PEG and LightSIDE, which rely on statistical techniques, and IEA, which is based on latent semantic analysis (LSA), most AWE programs evaluate texts using a branch of artificial intelligence called natural language processing (Grimes & Warschauer, 2010; Shermis,

2020). Natural language processing (NLP) mimics how humans naturally use language and develops computer programs capable of processing and understanding language in a human-like manner (Crossley, 2013). These programs can provide feedback on a wide range of aspects such as grammar, usage, style, mechanics, vocabulary usage, discourse structure, sentence variety, and discourse coherence quality (Deane, 2013; Landauer et al., 2003; Page, 2003).

AWE programs employing the latent semantic analysis LSA approach also have the capability to analyze the content of writing (Stevenson, 2016). This approach or method determines the similarity in meaning between words and passages through the analysis of large text corpora (Landauer et al., 1998). For instance, if an AWE tool is designed to evaluate scientific essays, it must first be trained on a dataset consisting of relevant scientific texts, such as textbooks or academic papers in that field. LSA helps break down textual information, allowing the AWE tool to understand both context and meaning, enabling it to assess how well a student's essay aligns with the expected content and structure of a given topic. However, LSA approach primarily focuses on the semantic appropriateness of the essay rather than on mechanical aspects like spelling and grammar. The Intelligent Essay Assessor (IEA) is an example of an AWE tool that utilizes the LSA approach.

Several applications and testing platforms have been developed to incorporate recent technology and research to facilitate the educational process. To aid in the evaluation of writing exams, well-known institutions such as ETS, Pearson, and Cambridge have developed AWE tools to serve their language centres. The first attempt to create an AWE tool was the Project Essay Grade (Page, 1966), which graded essays based on content and writing quality (Rudner & Gagne, 2001). Project Essay Grade (PEG) was developed by a network of universities in the United States to assess high school students' writing. Since then, other AWE programs have

been developed. Some of the most well-known and continuously evolving essay-marking programs include BETSY, the IntelliMetric program, the Intelligent Essay Assessor (IEA), and e-rater (Hutchison, 2007; Shermis, 2020).

Loncar et al. (2023) reviewed studies that investigated the use of technology-mediated feedback in L2 writing between 2015 and 2019. They found that a large number of studies investigate the use of automated writing evaluation (AWE), highlighting its affordances and emphasizing that, in general, both teachers and learners can access and use these tools comfortably. (Nunes et al., 2022) also reviewed studies conducted between 2000 and 2020 that examine the use of AWE tools in Grades 1–12. Their review aimed to identify the characteristics of various AWE tools and assess their impact on students' writing outcomes. The findings provided positive evidence supporting the usefulness of AWE systems for writing instruction and learning.

To illustrate the range of available AWE tools, Table 1 summarizes some features of AWE tools that generate formative feedback on students' essays, presented as examples to illustrate the variety of tools that exist and are actively researched. These programs are continually being developed and sponsored, with several (such as IntelliMetric, e-rater, IEA, and Critique) being used to evaluate students' writing in high-stakes tests (Warschauer & Ware, 2006; Nunes et al. 2022).

Table 1 *A Proposed Plan for Integrating AWE into EFL writing instruction.*

AWE tool	Scoring engine	Method	Feedback	Teacher role
My Access! Developer: Ventage learning	intellimetric	NLP	The system provides holistic scores and immediate feedback on specific traits: focus and purpose, content and development, organization, language use, grammar, and mechanics.	Teacher can comment and provide additional feedback
Criterion Developer: ETS	E-Rater	NLP	Wide range of individualized feedback on linguistic structure, style, organization, development of ideas, lexical complexity, prompt specific vocabulary usage.	Teacher can comment and provide additional feedback
Holt Online/ Essay Scoring/ Summary Street Developer: Pearson Knowledge Tecnology	IEA	LSA	Limited individualized feedback	NA
EssayCritic Developer: Intermedia	NA	LSA	Writers receive feedback as either "Praise" or "Critique", depending on the subthemes they covered. The	Teacher helps in providing samples to the system to be

			features measured are limited to the content and organization of short texts.	compared with students' writings.
NC Write PEG writing Developer: measurement incorporated	PEG	Statistical	A descriptive evaluation and feedback are provided on development of ideas, sentence structure, word choice, organization and style.	Teachers can provide in-text or summary comments.
Writing Roadmap Developer: McGraw-Hill	Bookette	NLP	The system provides immediate feedback on specific traits, highlights problematic sections, provides narrative comments. The feedback focuses on word choice, sentence structure, mechanics, organization, development.	NA

Note. IEA, intelligent essay assessor; PEG, project essay grade; LSA, latent semantic analysis; NLP, natural language processor; NA, not available.

Examples of AWE tools were added by the researcher. The table divisions and labels were adapted from Nunes et al. 2022.

3.4 The Effectiveness of AWE as a Scoring System

With the need to apply technologies in classrooms, AWE applications gained a lot of interest from researchers and educators. Several studies have found equality between human raters and AWE generated scores (Cohen et al., 2018; Duwairi, 2006; Hutchison, 2007).

In Cohen et al. (2018) study, 500 essays were randomly selected from the "Psychometric Entrance Test" which is a high-stake test done by higher education candidates as part of the university entrance requirements. The essays were then marked by 14 trained raters. All the raters followed the same rubric. For each essay the average of the 14 human scores was calculated by the researcher to represent the "true" score. These true scores were then compared to the automatically generated score and to a single (one of the 14 human raters that previously scored the texts) human rater's score. The results indicated that the automatically generated scores correlate with human raters scores in the same degree human scores correlate with each other, suggesting that the AWE tool scores just as well as human raters.

Along the same lines, Hutchison (2007), has also tried to examine the reliability of automated scores. His study was based in the UK. He collected 600 essays that were written by 11-year-olds on four topics: two narratives and two non-narratives. Then, they were scored by 3 human markers and compared to the automatically generated e-rater scores. The results show that the e-rater scores agree with human raters scores as nearly as human raters agree with each other. This aligns with Duwairi (2006), who compared automatic grading and human grading of university students and found that there was an agreement between the machine and human raters.

Powers et al. (2002) have also investigated if the e-rater scores were reliable. However, he tried to challenge the e-rater to find out what were the criteria that the system failed to evaluate. In the

study, professional writers were asked to write essays in a way to try to fool the AES program into awarding them higher than deserved or lower than deserved scores. Most of the chosen participants succeeded in receiving higher than deserved scores by the e-rater. There were discrepancies between human rater and the e-rater that reached in some cases 5 points out of 6. One of the participants wrote one paragraph and repeated it 37 times. He got the highest point from the e-rater (6) and the lowest point from the human graders (1). Another participant repeated the same paragraph by only rephrasing the first sentence of each paragraph. He also got the highest score from the e-rater. Only two participants managed to get lower than deserved points by the e-rater. That was by using metaphors and literary allusion and avoiding the use of keywords. That reveals that the e-rater has difficulties in evaluating intrinsic values. In spite of the fact that computers can simply miss out some features and abstract qualities of a written text (Hutchison, 2007), they do not have any problems when it comes to mechanics and organization.

Despite this, there are also factors that might affect the quality of human scoring. The first issue can be related to training and experience. In this context, Powers et al. (2015), compared automated scores with the grades given by one experienced, trained rater and another untrained rater. He found some discrepancies between the machine and the human grader when the grader is untrained and has less experience. This suggests that the agreement between the automated score and human scores relies on other factors, one of which is the grader training and experience. Secondly, Rudner (1992)found that even with sufficient instructor training, other variables such as background and experience of the rater can lead to slight but significant differences in grading. Third issue that might affect human scoring is the grader’s tiredness or fatigue (Bridgeman, 2013). He found that, in addition to the grader's experience, the grader's fatigue can also lead to fluctuations in scoring. Rater can become tired by the end of a long day

of scoring essays or after a long week of testing session. Tiredness may make him/her look at the external features of the essay rather than the quality of the writing and the argument. In contrast, AWE scoring systems are not affected by such factors, which make them more consistent in awarding scores to a large number of students.

A final significant issue related to the effectiveness of AWE scoring is time consumption. Rudner and Gagne (2001) asserted that assessing students' writing is one of the most expensive and time-consuming activities in assessment programs. Marking papers consumes long hours and graders must be punctual and meet strict deadlines, which is crucial for maintaining the smooth flow of the educational process. For example, in the study of Cohen et al. (2018), some human graders were excluded from participation because they failed to meet the deadline provided to submit corrected papers. Computer scoring, hence, offers a potential solution as it is faster, reduces costs, increases accuracy and eliminates concerns about graders' consistency and fatigue (Rudner & Gagne, 2001). However, a limitation of relying solely on e-rating is its inability to assess the intrinsic value of writing. The quality and coherence of ideas remain complex elements that machines cannot yet accurately measure.

In light of these benefits and limitations, several empirical studies have examined the effectiveness of AWE feedback in improving writing quality, exploring how automated feedback can complement human evaluation in enhancing students' writing performance.

3.5 The Impact of AWE Feedback on Writing Quality

Besides automated essay scoring, another widely investigated feature provided by most AWE systems is the automated feedback they generate. Hyland and Hyland (2006b) argue that AWE feedback can be a good example of computer-mediated communication (CMC), as it can

empower students, make writing classes more collaborative, and thus elevate the social aspect of learning.

Consequently, there has been growing interest in recent years in the use of AWE systems to offer formative feedback in writing classrooms (Stevenson & Phakiti, 2014). AWE systems are able to generate feedback on different aspects of writing, including language use and mechanics (e.g., Criterion, MY Access, Grammarly), content and style (e.g., Criterion, WriteToLearn), or rhetorical quality (e.g., Intelligent Academic Discourse Evaluator (IADE)). Considering their increasingly popular applications for formative assessment in writing classrooms, a number of studies have analyzed them, compared them to other types of feedback, and examined their potential as learning tools in writing instruction (Attali, 2004; Dikli & Bleyle, 2014; Van Der Kleij et al., 2012; Wilson & Czik, 2016).

Stevenson (2016) critically synthesized research on the use of automated writing evaluation (AWE) in writing instruction. Her synthesis focused on three key areas: the purpose of using AWE, its integration into teaching practices, and teachers' and students' perceptions of AWE use. Regarding the purpose of AWE, Stevenson found that it was primarily used to save teachers time, promote learner autonomy, and support the development of writing processes. Additionally, her synthesis identified less frequently examined purposes in AWE research, such as promoting social interaction and developing content knowledge. In terms of AWE integration, most studies examined the use of AWE to augment teacher feedback through embedding it in classroom instruction, assessing students writing, preparing students for exam, and facilitating collaborative learning with peers. Other studies looked at the role of teachers in scaffolding the use of AWE. The third area of investigation examined the usefulness of AWE. Overall, both students and teachers responded more positively to AWE-generated error feedback than to other

aspects of feedback, such as scoring or holistic evaluation. However, several limitations emerged, including technical issues, concerns about scoring accuracy, difficulties in interpreting AWE feedback, and students' reluctance to engage in revision. As a potential solution, it is suggested that AWE focus on providing error correction in the initial phases of writing a text, freeing the teacher up to concentrate on higher-level meaning-oriented, genre-oriented and audience-oriented aspects of writing, or using AWE to increase critical awareness of what writing feedback involves (Stevenson, 2016; p.12).

In a more recent research synthesis, Karatay and Karatay (2024), critically investigated literature on student engagement with AWE in L2 classrooms and the impact of using automated feedback on writing. The synthesis examined studies published between 2013 and 2021 in the context of second and foreign language. The findings revealed that AWE feedback positively contributed to students' writing accuracy across diverse contexts and tools. In addition to reducing error rate, the synthesis highlighted that AWE tools consistently create learning opportunities beyond specific writing tasks and provide suggestions that expand learners' linguistic knowledge. Engaging with AWE helped learners to become more independent and increases learner autonomy. The study showed that revision practice positively influenced students' engagement with the text and also improved teacher feedback practices. In conclusion, the synthesis of the reviewed studies emphasizes the importance of integrating AWE tools with teacher feedback and highlights the need for a balanced approach to using both types of feedback in support of L2 writing instruction.

Generally, AWE applications are valued for their ability to provide instant, individualized, and specific feedback, which can positively influence learners' writing accuracy. According to Bridgeman (2013), AWE systems offer organized, qualitative feedback, particularly related to

the structure and form of writing. In writing classrooms, students often require not only more examples and practice but also continuous feedback to refine their writing skills. AWE systems cater to this need by enabling learners to submit their writing, receive immediate feedback, revise their work, and then resubmit for further evaluation. This cycle of feedback and revision fosters a more interactive and dynamic learning environment, where students have multiple opportunities to improve their writing through continuous refinement. Furthermore, by automating this process, AWE systems can alleviate some of the workload for instructors while still providing students with the detailed, formative feedback they need to progress. This makes AWE tools particularly effective in large classes or for students who need frequent, timely responses to their work.

Regarding the impact of AWE on writing accuracy, Barrot (2023) found that when learners engaged in a cycle of revising and resubmitting multiple writing drafts, their writing accuracy improved, and the number of error types decreased. In her quasi-experimental study, she examined the effect of using Grammarly, an automated writing evaluation (AWE) tool, on students' writing accuracy. Sixty-five ESL students from a private university in the Philippines participated, submitting argumentative essays weekly over 14 weeks. All participants were pre-assessed using an institutional admission test, confirming intermediate proficiency (B1 level, CEFR). Both groups attended the same classroom instruction but only the experimental group used Grammarly as their sole feedback source, while the control group received no feedback. To assess participants' progress, a pretest and a posttest in writing were administered. The results showed a decrease in the number of errors for both groups and an increase in posttest scores. However, only the experimental group achieved statistically significant results with a large effect size.

These results align with research advocating for timely, automated feedback (Stevenson & Phakiti, 2014), yet Barrot's study raises critical concerns. While Grammarly provided grammatical and mechanical corrections, the exclusion of human feedback might limit the depth of revisions, particularly concerning content, argument development, and critical analysis—areas where human feedback traditionally excels (McNamara et al., 2015).

Moreover, the study's limited sample size and single context (a private university in the Philippines) may restrict the generalizability of the findings. Future research could explore whether these results are held in different educational settings or with larger, more diverse populations. Additionally, while Barrot demonstrates the effectiveness of AWE in reducing error types, she does not examine whether learners are internalizing these corrections in a way that improves their long-term writing competence beyond mechanical accuracy.

Hung et al. (2024) also advocate for the use of AWE tools in the revision process, emphasizing their potential to improve writing proficiency. Their study suggests that integrating AWE software into revision activities can lead to measurable improvements in writing skills. In their research, 59 EFL students enrolled in an introductory English course were divided into two groups. The experimental group (30 students) revised their work based on AWE-generated feedback, while the control group (29 students) followed a traditional drafting process for revisions. Pretest and posttest assessments were conducted to evaluate the effectiveness of AWE-assisted revision. The results of the t-test analysis revealed a statistically significant improvement in the writing proficiency of the experimental group compared to the control group. Notably, students in the experimental group demonstrated a substantial reduction in error rates and a higher word count in their revised drafts, indicating a positive impact of AWE-assisted feedback on their writing development.

While the study supports the use of AWE to improve EFL students' writing proficiency, several limitations must be considered. First, there was no clear information regarding the feedback process used for the control group. This raises the possibility that the observed improvement may be attributed to training effects rather than the impact of AWE itself. Additionally, the study did not specify which writing features showed improvement, making it difficult to determine whether AWE primarily influenced grammar, spelling, vocabulary use, or overall content and organization.

Moreover, assessing word count as a measure of writing improvement may not be reliable. Previous studies have identified limitations in AWE systems, as some students can manipulate the system by repeating portions of their writing to achieve higher scores (Power et al., 2002). Wilson and Czik (2016) also found that AWE can "be easily fooled to assign high scores to essays which are long, syntactically complex, and replete with complex vocabulary" (Wilson & Czik, 2016, p. 95). This highlights the critical role of instructors in ensuring that students engage meaningfully with AWE-generated feedback and that revisions reflect genuine improvements rather than score-boosting strategies.

To investigate further, Ranalli (2018), examined the impact of generic versus specific feedback within the AWE software. He compared participants' writing accuracy by analyzing the number of error corrections made when they received specific feedback versus generic feedback. In his study, eighty-two ESL students participated: 36 from two sections of a lower-level course (an academic writing course focusing on sentence and paragraph structure) and 46 from five sections of an upper-level course (an academic writing course focusing on essay-length assignments and practice in writing processes such as pre-writing, revising, and editing). Participants in both groups used Criterion to receive automated feedback. The feedback was

classified as either generic (a general solution offered to fix the error, e.g., "proofread this!") or specific (when a part of the text was flagged and a specific solution was offered, e.g., "You have used 'a' in this sentence. You may need to use 'an' instead!"). A corpus of Criterion data was compiled from all drafts of writing submitted to the system. Error reports were generated, saved, and analyzed to quantify the instances of each error type flagged by the system. The findings showed that generic feedback resulted in fewer successful error corrections than specific feedback. Additionally, course level was not found to be a significant factor in most analyses.

Similar to Barrot (2023), a limitation of Ranalli's study is its narrow focus on certain writing issues (i.e., grammar, usage, and mechanics). While important, these aspects represent only a portion of what constitutes writing proficiency, especially in academic settings. The study does not address other writing areas such as content development, argumentation, or organizational structure, which are also critical for writing improvement.

Kleij et al. (2012) compared students' learning outcomes and attitudes after receiving immediate and delayed elaborate feedback. In their experimental study, 152 first-year university students participated under three different feedback conditions. The first group received immediate computer-based feedback along with elaborate feedback, which provided detailed explanations of the correct answers. The second group received delayed computer-based feedback and delayed elaborated feedback. The third group received delayed computer-based feedback with only knowledge of results (i.e., whether the answer was correct or incorrect, without further explanation). The posttest summative assessment results indicated no significant differences in learning outcomes among the three groups. However, survey results revealed that students expressed a more positive attitude toward the immediate elaborated feedback, which was provided by the AWE tool.

It is worth mentioning that the findings of Kleij et al. (2012) reinforce the advantages of timely feedback provided by AWE software to improve learning. The students who participated in the experiment preferred immediate elaborate feedback, even though no significant differences in performance outcomes were observed. However, the study's short-term scope, lack of revision analysis, and absence of teacher feedback comparison highlight areas for further investigation. Future research should explore long-term learning effects, student engagement with feedback, and hybrid feedback approaches to provide a more comprehensive understanding of how AWE tools can best support L2 writing instruction.

Other studies have also shown that automated feedback has a positive impact on reflective learning skills focusing on the semantic content, which may contribute to improving writing quality. Cheng (2017) and Cheng et al. (2017), found that automated feedback can effectively promote students' reflective learning abilities, showing a positive correlation between reflective learning skills and essay writing proficiency. Cheng (2017) used a self-developed AWE software, a web-based automatic classification system, to generate immediate and individualized feedback on students' reflective journals. The system relied on Latent Semantic Analysis (LSA), focusing on content rather than language use and mechanics (Landauer et. al., 1998; Stevenson, 2016). The study aimed to investigate whether the use of this software could positively impact participants' reflective learning and, consequently, improve their writing quality. Results revealed a significant increase in writing scores among the experimental group from the first to the second journal and from the second to the third journal. In contrast, there were no significant differences in the control group's scores.

While these findings support the argument that AWE tools can foster reflective learning and writing quality, it is essential to consider the limitations. The study's reliance on a self-

developed AWE tool, tailored specifically to reflective journals, raises questions about the generalizability of these results to other types of writing tasks. Additionally, the exclusive focus on content feedback (using LSA) might neglect other crucial aspects of writing, such as syntax and grammar, which also influence writing quality. Further studies are necessary to explore how automated feedback impacts different dimensions of writing, as well as to assess its effectiveness across diverse genres and educational settings.

Regarding this concern of investigating the impact of AWE of different genres, Zhu et al. (2020) examined the use of AWE to improve scientific argumentation writing. In the study they compare the impact of generic feedback (context-independent) vs. contextualized feedback (context-dependent) on writing scores. They also explore how student revisions enabled by the formative feedback system correlate with student performance and learning gains. The dataset included 374 students from 7^{th}-12^{th} grade across 22 classes. They were instructed by eight teachers from eight different schools across the United States. Students in each class were randomly assigned to either the generic feedback condition or the contextualized feedback condition. Making revisions was voluntary after receiving the automated feedback. Students could make as many revisions as they wanted. After each revision they also received updated automated feedback and scores. The findings revealed that students' revisions positively impacted students' scores and contextualized feedback was more effective than generic feedback in assisting learning.

These results are in line with previous studies that had examined the correlation between revision frequency and students' writing proficiency (Barrot 2023; Hung et. al. 2024). With regard to comparisons between specific and generic feedback, the results also align with Ranalli (2018), who confirmed the effectiveness of content-specific feedback over generic feedback.

Although this study contributes to the literature on the impact of AWE feedback on a specific writing genre (argumentative writing), further research is needed to compare different genres to determine whether the effectiveness of automated feedback varies across writing types. Another limitation of this study is its focus on L1 students' writing, which raises questions about the generalizability of the findings. These results may differ among participant groups with varying levels of English proficiency, highlighting the need for further exploration in L2 writing contexts. Additionally, the effectiveness of AWE could also be tested in comparison to other types of feedback (teacher, peers) to assess its relative impact on writing development.

3.6 Challenges and Advantages in AWE Implementation

While research generally supports the use of AWE tools in writing classrooms, some studies raise concerns about their practical effectiveness. A significant issue is learners' ability to interpret and utilize the system effectively. Some studies suggest that AWE may become more of a burden than a facilitator if students struggle to understand the feedback (Hyland & Hyland, 2006a; Ranalli, 2018). For instance, Hyland and Hyland (2006a) argued that students can become overwhelmed by the quantity and complexity of comments, making efficient revision difficult. Similarly, Ranalli (2018) observed that unclear feedback negatively affected students' willingness and ability to revise their work. In his study, some participants chose to delete parts of incorrect sentences rather than engage with the feedback and revise their work. He attributed this behavior to the feedback's lack of clarity, which made it difficult for students to act on. Consequently, Ranalli recommended that AWE developers prioritize explicitness in the feedback provided to ensure better outcomes. A potential solution to this issue is for teachers to offer guidance by training students on how to effectively respond to vague or ambiguous feedback.

In contrast, some studies suggest that AWE can be effective even for low-proficiency students. For example, Wilson (2017) explored the feasibility of using an AWE tool with students with difficulties (SWD). He discovered that SWD were able to close the gap in writing quality between themselves and typically developing students after five revisions of their drafts. His study demonstrated a positive association between automated feedback and improvements in writing quality for SWD. It also highlighted the importance of having adequate technological resources to maximize these improvements. Notably, it confirmed that even SWD could effectively use the tool to improve their writing.

Attali (2004) and Roscoe et al. (2014) also examined the effectiveness and usability of automated feedback. Attali (2004) explored the effect of *Criterion's* automated feedback on students' essay writing. In his study, he collected data from students' writing texts in grades 6 through 12 during the 2002–2003 school year. He compared the essay texts of first and last submission. He also analysed the automated scores generated, feedback reports, number of submissions, and the grade for which the essay was designed. In contrast to previous concerns, the results showed that participants were able to effectively respond to the automated feedback provided. They succeeded in reducing their error rates by approximately 25% and increased the number of main points and supporting ideas, which resulted in longer essay lengths.

While Attali (2004) provided important evidence that students can respond to automated feedback, the study had several methodological limitations. First, there was no information about participants proficiency level or writing competence background. In addition, the focus of the study was on short-term progress without assessing whether the observed improvements were sustained over time. Furthermore, the lack of a control group limited the study's ability to confirm that the effects of automated feedback were solely responsible for the improvements.

Without a comparison to teacher feedback or a no-feedback condition, it is difficult to draw strong conclusions about the effectiveness of AWE.

Roscoe et al. (2014) also reported that students successfully engaged with the AWE tool, finding it informative, valuable, and enjoyable. However, similar to Attali (2004), the study focused on the usability of Writing Pal (W-Pal) as an automated writing evaluation system without comparing it to other types of feedback. Typically, studies of this nature are designed for a purpose to improve AWE system rather than to validate the effectiveness of its feedback in improving writing outcomes.

Lastly, due to inconsistent findings and methodological limitations, further research is needed to determine whether EFL learners can effectively use AWE systems independently and comprehend the feedback provided without teacher assistance. Moreover, future studies should compare automated feedback with teacher feedback to assess their relative effectiveness in supporting writing development.

3.7 AWE Feedback vs. Teacher Written Feedback

To examine the impact of using AWE-generated feedback, numerous studies aimed at comparing it to the traditional teacher feedback which is widely used in writing classrooms (Chandler, 2003; Dikli, 2010; Dikli and Bleyle, 2014).

Dikli (2010), compared feedback from My Access! (an AWE tool) to teacher feedback and noted significant differences. Teacher-written feedback was concise, focused, and specific, while the My Access! feedback was generic, overly lengthy, and sometimes redundant. It is important to note, however, that this critique of My Access! maybe system-specific and should not be generalized to all AWE tools, as the quality of automated feedback can vary widely depending on the system's design and algorithms.

In contrast, Chandler (2003), found that instructors often provide general comments and praise for students' efforts. He argues that while this type of feedback may encourage students, they benefit more from specific, detailed feedback. However, providing such individualized feedback on a regular basis is a challenge for teachers, especially with larger class sizes. These contradictory findings highlight the need for a systematic examination of the differences between teacher feedback and automated feedback. Further research should explore what the best method of integrating these feedback types is, to improve writing instruction and maximize student learning outcomes.

Building on these observations, other studies delved deeper into the content and quality of AWE-generated feedback compared to teacher feedback, while also exploring students' perceptions of these tools. For instance, Dikli and Bleyle (2014), compared *Criterion's* (AWE tool) feedback with instructor's feedback. They found that the instructor provides more accurate and better-quality feedback. He then used a survey to further explore participants' perceptions. Results showed that some of the participants recognized the weaknesses in *Criterion's* ability to provide feedback. Surprisingly, most of the participants seemed to trust *Criterion*'s feedback and felt motivated to use the program to revise their writing. Students' trust in AWE feedback might stem from the immediacy and availability of the tool rather than the quality of the feedback itself. This could lead to overconfidence in flawed feedback, potentially limiting the development of critical revision skills. Furthermore, the study's focus on student perception, while valuable, leaves open the question of whether Criterion's feedback meaningfully contributes to writing improvement. Future research should critically assess whether such tools truly improve learning outcomes or merely provide superficial revisions that students find reassuring.

In addition to earlier investigations, Weigle (2013), suggests that teachers can use automated writing evaluation to assess the structure issues of writing such as grammar errors and mechanics. This allows them to focus more on other problems in writing such as fluency and content. Along the same line, Ware (2011), demonstrates that automated writing evaluation should be used for writing assistance rather than for writing assessment.

3.8 AWE and Teacher Feedback: A Combined Approach

Literature on the use of automated writing evaluation (AWE) has recommended that the optimal feedback approach is to integrate AWE with teacher feedback (Ware, 2011; Weigle, 2013). Additionally, Karatay and Karatay (2024), in their recent review of AWE research, emphasize the importance of combining AWE tools with teacher feedback. Their findings highlight the need for a balanced approach that leverages both automated and human feedback to effectively support L2 writing instruction. Thus, a hybrid feedback system is thought to have the potential to reduce issues related to teacher only or automated only feedback.

Wilson and Czik (2016) argue that AWE systems are expected to help teachers provide more high-level feedback, expedite the feedback process, and support improvements in students' writing motivation and writing quality. In their quasi-experimental study, four eighth-grade English Language Arts (ELA) classes received feedback from both their teacher and the PEG Writing system, while another four classes received only teacher feedback through Google Docs. The results showed that although teachers provided a similar amount of feedback across both groups, those in the PEG + Teacher Feedback condition received proportionately more feedback on higher-level writing skills (e.g., ideas and elaboration, organization, and style). Teachers also reported that PEG Writing helped them save between one-third to half of the time normally required to provide feedback compared to when they were the sole source of feedback in the

Google Docs condition. While students in the combined feedback group demonstrated increased persistence in writing tasks, no significant differences were found in the final-draft writing quality between the two groups. This study confirms the previous findings and highlights the potential time-saving benefits of AWE systems for teachers and suggests that AWE can promote the focus on higher-order writing skills, although the impact on final writing quality remains inconclusive.

Grimes and Warschauer (2010) also propose that a mindful use of AWE can encourage L2 learners to write, revise, and practice more frequently. In their multi-site case study, they examined how 'My Access,' an AWE tool, was implemented in eight U.S. middle schools over a three-year period. The data set included classroom observations, interviews with principals and teachers, and surveys of both teachers and students. Additionally, they collected AWE-generated reports to analyze the average number of essays written per student per year and the average number of revisions per essay. The study found that when teachers integrated AWE as planned, it simplified classroom management, allowing teachers to feel more relaxed and focused. Students also demonstrated positive attitude toward using AWE software in class. Observations suggested that students were more engaged with writing tasks while using the system compared to traditional pen-and-paper methods.

Along the same line, Link et al. (2020), propose that the most effective use of automated written feedback in the classroom is as a "complementary source" rather than a standalone tool. In their study, they compared two groups: one group received feedback solely from their teacher, addressing both higher-level (HL) writing skills (ideas, elaboration, organization, and style) and lower-level (LL) writing skills (spelling, capitalization, punctuation, structure, grammar, and word choice), while the other group received automated written feedback on (LL) writing skills,

along with teacher feedback focused on (HL) writing skills. The study involved 32 undergraduate English majors. The participants were randomly assigned to either AWE + Teacher feedback group or Teacher feedback group.

The study design included a pretest, posttest, and delayed posttest to measure changes in writing performance. The findings revealed that using AWE as a complement to teacher feedback did not significantly affect the amount of (HL) feedback provided by the teacher. However, the teacher who did not use AWE tended to offer more (LL) feedback than AWE alone provided. Furthermore, students were more likely to revise teacher-provided LL feedback compared to similar feedback from the automated system. Notably, the study also found that students who had access to AWE retained their accuracy improvements from pretest to delayed posttest, while those who did not use AWE showed less retention. These findings underscore the value of AWE as a supplement to teacher feedback, particularly in promoting long-term retention of writing accuracy. However, its impact on the quantity and type of feedback provided may vary depending on how it is integrated into classroom instruction. Additionally, it is important to examine students' writing accuracy across different aspects of writing, as previous research has indicated that the success of students' revisions is often closely tied to their overall writing accuracy.

Palermo and Thomson (2018) assessed the impact of integrating Self-Regulated Strategy Development (SRSD) and teacher instruction with the AWE tool NC Write on students' argumentative writing performance. A total of 829 middle school students participated in the study, which randomly assigned them to one of three conditions: teacher instruction only, NC Write + teacher instruction, and NC Write + SRSD. The results revealed that students in the NC Write + traditional instruction condition produced higher-quality essays than the comparison

group (teacher instruction only) at the posttest. Furthermore, students in the NC Write + SRSD condition not only wrote essays of superior quality but also produced longer texts and included more essential elements of argumentative writing compared to students in the other two groups. These findings suggest that combining AWE systems with highly effective writing strategies like SRSD can significantly improve students' writing quality.

While the study demonstrates the potential benefits of combining AWE with structured writing strategies, it is important to critically assess whether these findings are generalizable across different student populations and educational contexts. The use of middle school students, for instance, may limit applicability to other age groups or learning environments. Additionally, while NC Write + SRSD showed the most promise, the extent to which these tools reduce teacher workload or foster student independence remains unclear. Moreover, future research is needed to study whether AWE is associated with improvements of writing quality when students compose independently.

Sari and Han (2024) investigate the impact of a combined feedback condition (teacher + automated feedback) in EFL context. Using a quasi-experimental pretest-posttest control group design, they examined the effects of automated feedback on sentence-level errors alongside teacher feedback on content and organization. This hybrid feedback approach was compared to a traditional teacher-only feedback condition to assess its effectiveness. Two intact classes were randomly assigned to the experimental and control groups, and the study was conducted over a 16-week semester. Writing test results were analyzed to determine whether the integration of automated and teacher feedback led to greater improvements in EFL students' writing performance compared to conventional teacher-only feedback. A questionnaire and focus group interviews were also used to explore students' experience. The results show that the use of

combined automated-teacher feedback was more effective than conventional teacher-only feedback in enhancing the students' writing performance. The qualitative data also shows that students had favourable opinions of their experiences with receiving automated and teacher comments together.

Sari and Han (2024) make a valuable contribution to AWE research by examining the benefits of a hybrid feedback approach. Their findings support the idea that combining AWE and teacher feedback could develop writing performance, particularly by leveraging AWE for sentence-level errors and teachers for higher-order concerns. However, the study has some methodological and theoretical limitations, including lack of detail on feedback implementation and limited assessment of writing feature improvements. The study does not clarify which specific aspects of writing improved the most: sentence-level corrections (grammar, mechanics) or content-related revisions. Thus, a more detailed breakdown of writing improvements would strengthen the study's conclusions. Furthermore, it is worth considering the long-term development. Addressing these gaps would provide a more comprehensive understanding of integrating AWE into EFL writing instruction.

Contrary to Sari and Han (2024) findings, Fan (2023) found that students who used a hybrid feedback condition did not outperform the control group who received teacher feedback only. In his study, he examined the impact of using a hybrid feedback condition on EFL students' writing quality. The participants were 67 students divided into experiment and control group. The treatment group received teacher feedback on content and organization and then used Grammarly to receive feedback on spelling, grammar, and punctuation while the control group received only traditional teacher feedback. The study used CAF measure (Complexity, Accuracy, and Fluency) to compare between groups. The results of the posttest revealed that the students

from the treatment group did not significantly outperform the students from the comparison group in syntactic and lexical complexity, accuracy, and fluency. Although the study contributes to the importance of implementing AWE, the findings have several limitations. The study has a short-term duration, a longitudinal study could help determine whether hybrid feedback has a delayed effect on writing proficiency. These results might also be linked to the software used. The use of Grammarly is widely researched, however, its effectiveness was mainly linked to corrective feedback.

The conflicting findings between Fan (2023) and Sari and Han (2024) indicate that the effectiveness of hybrid feedback may depend on contextual factors, including learner proficiency, feedback implementation, and AWE software used. These findings emphasize the need for continued exploration of how best to integrate AWE and teacher feedback to improve L2 writing instruction.

3.9 Summary

The body of literature reviewed demonstrates the evolving role of automated writing evaluation (AWE) systems in writing instruction. While several studies support the potential benefits of AWE tools, particularly as a complement to teacher feedback, there are still critical areas where further research is necessary to draw more definitive conclusions.

One recurring theme is the challenge of balancing AWE and teacher feedback. Chandler (2003) and Dikli (2010), highlight the limitations of AWE systems in providing the nuanced, detailed feedback that students often need, particularly in terms of content and higher-level writing skills. In contrast, research by Wilson and Czik (2016) and Link et al. (2020), suggests that combining AWE tools with teacher feedback can help teachers focus on higher-level feedback while still benefiting from the time-saving features of automated systems. However, the

literature often focuses on a comparison between just two groups: AWE-only feedback versus teacher-only feedback or AWE + teacher feedback. This creates a gap in understanding how the three distinct conditions: AWE-only, teacher-only, and AWE + teacher feedback, compare in terms of their impact on student writing performance. Moreover, recent studies that investigate the implementation of hybrid feedback condition (automated feedback + teacher feedback) show contradictory findings (Fan, 2023; Sari & Han, 2024).

In addition, most studies to date have focused on general writing improvements or specific feedback types (e.g., high-level versus low-level feedback), without delving deeply into how AWE and teacher feedback might differently affect various aspects of writing, such as: content, organization, vocabulary use, language use, and mechanics. As Weigle (2013) and Ware (2011) suggest, AWE systems may be more effective when used for mechanical aspects of writing, while teachers are better suited to address content and fluency issues. Nevertheless, further research is needed to systematically compare how AWE and teacher feedback influence these distinct writing components across different instructional settings.

Overall, the literature does not point in one unified direction but suggests a more nuanced view. Some studies report that AWE tools can perform as reliably as human raters (Cohen et al., 2018), yet both AWE and human scoring face persistent challenges. A more balanced interpretation emerging from recent studies is that AWE systems are particularly effective for lower-level aspects of writing such as mechanics, grammar, and surface errors (Weigle, 2013; Ware, 2011), whereas teacher feedback is more valuable for higher-level concerns such as content, argumentation, and organization (Chandler, 2003; Dikli, 2010). This has led to calls for hybrid approaches where AWE handles mechanical issues, enabling teachers to focus on meaning and discourse (Wilson & Czik, 2016; Link et al., 2020). However, contradictory

findings regarding hybrid conditions (Fan, 2023; Sari & Han, 2024) show that this division of labor is not yet fully confirmed, leaving open the need for studies that systematically compare AWE, teacher, and hybrid feedback.

Another area that remains underexplored in the literature is the role of genre in the effectiveness of automated writing evaluation (AWE) tools. Much of the existing research focuses on general writing development without accounting for how different genres (e.g., expository vs. persuasive) may interact with the type of feedback provided. However, genre-specific writing tasks involve distinct cognitive and rhetorical demands, which may influence how learners process and benefit from feedback.

Drawing on the cognition hypothesis and the concept of task complexity, persuasive writing is typically considered more cognitively demanding than expository writing, as it requires critical thinking, the synthesis of multiple viewpoints, and the construction of logical arguments. This genre-based variation provides a valuable framework for investigating how feedback interacts with the cognitive demands of different writing tasks.

Given these considerations, there is a clear need for studies that assess the impact of AWE tools across various genres to determine whether their integration into writing instruction should be tailored to specific writing tasks. In response to this gap, the current study also contributes to research on task complexity by examining which type of feedback—teacher, automated, or hybrid—is most effective in supporting learners across tasks of differing complexity. Specifically, it explores whether the effectiveness of feedback varies by genre and whether more cognitively demanding tasks benefit more from certain types of feedback.

In summary, the current thesis aims to address several key gaps in the literature. First, it

compares three feedback conditions: (1) automated feedback only, (2) teacher feedback only, and (3) hybrid feedback (automated + teacher). By doing so, it aims to provide a more comprehensive understanding of how different feedback conditions may impact on writing improvement. Second, the study examines specific components of writing—text content, text organization, vocabulary use, language use, and mechanics—to determine which aspects benefit most from the feedback provided. Lastly, the research explores how feedback effectiveness may differ across writing genres (expository and persuasive), thereby contributing to a more nuanced understanding of how AWE systems can be integrated into diverse writing instruction contexts. Together, these aims not only address current gaps in the literature but also offer practical implications for educators seeking to optimize the use of AWE tools in EFL writing classrooms.

4. Methodology

4.1 Overview

This chapter describes the research methodology and data collection methods employed during the treatment. It begins by presenting the proposed research questions that guided the investigation. Following this, information is provided regarding the participants and the study's context. The chapter then explains the research design, data collection procedure, and data analysis processes. The pilot study procedure conducted prior to the main study is outlined, and issues associated with validity and reliability are discussed. Finally, the chapter accounts for the ethical issues considered both before and during the intervention.

4.2 Research Questions

The research questions driving the present project are as follows:

RQ1: To what extent do three different feedback conditions (teacher only, automated only, hybrid feedback) affect overall L2 writing production?

RQ2: To what extent do three different feedback conditions (teacher only, automated only, hybrid feedback) affect the different components of L2 writing examined: text content, text organisation, vocabulary use, language use and mechanics?

RQ3: Do the effects of three different types of feedback on writing differ depending on the genre of writing?

4.3 Research Design

The current study explores the effect of employing various feedback conditions (teacher feedback, automated feedback, hybrid feedback) on EFL students' written texts' quality. The research questions are causal questions because they seek to evaluate the relative effectiveness

of various feedback conditions interventions. Following the experimental design is the best method to answer such causal research questions properly. O'Leary (2014) demonstrated that carrying out an experimental study is thought to be the most effective way to conduct a thorough search of cause and effect which is the main aim of the current study. One of the best practices in an experimental design is the randomisation of the sample because it enables the findings to be generalised (Flick, 2020). However, in the educational context it would probably disturb classroom learning (Creswell & Guetterman, 2019). Similarly, Dörnyei (2007) argued that it is difficult to adopt a pure experiment design in an instructional setting. According to Lantolf (1999), experimental and quasi-experimental studies are highly relevant to classroom-based research, as they reflect authentic classroom practices. Therefore, the current study adopts a quasi-experimental design which is appropriate to address the stated research questions. It is similar to the pure experiment except for the randomisation of the sample (Cohen et al., 2018). Some educators even considered it to be an alternative and as good as a true experiment, more powerful than other designs and capable of establishing what is needed to support causal conclusions (Shadish et al., 2002). The current study compared three groups under different feedback conditions (see Figure 2). The groups were randomly allocated to one of the three conditions to provide protection against biased differencing between the groups (Gorard, 2013). The three groups were also considered equal in terms of the proficiency level, age, and number of years of English education. To ensure that the learners have the same level of English language proficiency at the start of intervention, prior to the treatment, all participants completed a paper Oxford Placement Test (OPT) (http:// www. oxfordenglishtesting.com/).

4.4 Participants and Setting

4.4.1 Students

The current study was carried out using a sample of 74 Saudi female students with an average age of 19 years. Their first language (L1) is Arabic and they were studying English as a foreign language as part of their foundation year at the University of Jeddah, Saudi Arabia. The foundation year prepares undergraduate students to major in a variety of subject areas. The writing instruction is part of an intensive English language course which focuses on developing the four skills of reading, writing, listening and speaking, as well as developing critical thinking skills and presentation skills. The students in the foundation year at University of Jeddah are assigned to different levels of proficiency in English according to their performance in high school exit exams. Then they proceed to the next level after passing the final exam at the current level. The participants in the current study were at the pre-intermediate level which correlates with B1 in the Common European Framework of Reference for Languages (CEFR). The participants usually take three sessions of English language classes daily with four sessions of writing practice per week. Each session lasts for 50 minutes.

The investigation took place during the first term of 2022/23, which spans 12 weeks. Sampling procedure utilized convenience sampling, with three intact classes, each with a total of approximately 30 students, being chosen for participation. The three intact classes were selected to participate based on their assigned schedules, following arrangements made with the registration office, the vice dean, and the scientific research unit of the University of Jeddah. Each class had a total of approximately 30 students.

Within each class, participants were randomly assigned to one of the three sub-groups based on the type of feedback they would receive during the intervention: automated feedback

(AF), teacher feedback (TF), and combined automated and teacher feedback (ATF), (see Figure 2). Random assignment was carried out using a the students University ID numbers to ensure equal representation of feedback types across sub-groups, thereby minimizing bias and maintaining a balanced study design.

To address potential practical challenges, several measures were taken. First, all feedback was provided in written form only, which prevented students from observing or comparing the type of feedback received by their peers. Second, teaching content and instructional procedures were kept consistent across all classes, with only the mode of feedback differing according to group assignment. Third, students were informed that they would receive different types of feedback as part of the study but were not made aware of which group their peers belonged to, further reducing the possibility of influence across groups.

In total, 74 participants completed the treatment sessions and the immediate posttest. The TF group had 25 participants, the AF group had 25 participants, and the ATF group had 24 participants. However, at the delayed posttest time, the number of participants had significantly decreased. Only 32 participants completed the delayed posttest: 12 from the TF group, 10 from the AF group, and 10 from the ATF group.

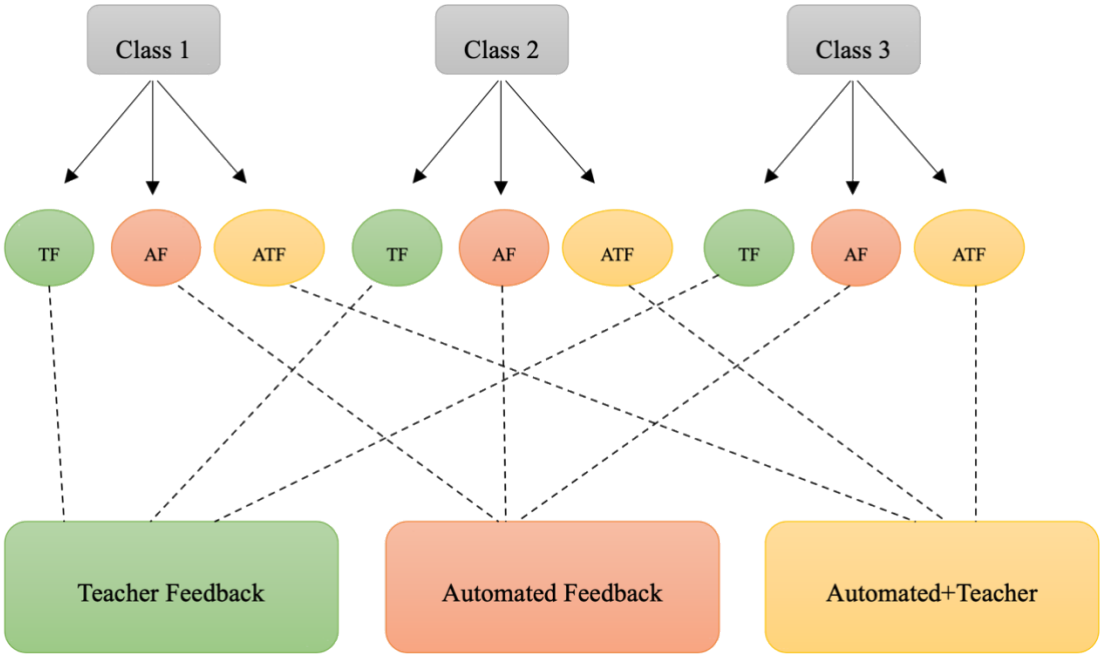


Figure 2 A Flowchart of Groups Assignment Process.

The students typically spend approximately 4 hours per week focusing on their writing skills. Their writing practice include the following: 1) writing straightforward connected texts on a range of familiar subjects, following standard layout and paragraphing conventions; 2) writing straightforward, detailed descriptions on a range of familiar subjects; 3) narrating a story; 4) writing short, simple essays on topics of interest; 5) summarising; 6) giving opinions; 7) writing reports regarding factual information on familiar topics. The writing composition practice usually involves students to write short paragraphs of 150-175 words and then elaborate and write longer essays of 250-300 words on various topics in different genres (e.g., descriptive, letter, narrative, report, and opinion). The procedure of writing practice follows the writing approach which involves drafting, receiving teacher feedback, re-drafting and submitting the

final draft for grading. The topics for writing practice are the same for all classes and are provided from the curriculum unit of the English Language Institute at the University of Jeddah.

The participants at the time of the intervention had only practiced writing a letter which is different from the genres chosen in the present study. This helps to avoid the effect of learning, an extraneous factor that might affect the validity and reliability of the research outcome (See section 4.9).

4.4.2 Teachers

A total of six teachers participated in the study. They were all master's degree holders and had a similar number of years of experience (8+) teaching English at the University of Jeddah.

Three teachers participated in the treatment sessions and were responsible for instructing the students and giving them feedback on writing when required (depending on the group they were assigned to). To control for variation between the instructors in the amount, content, form, and frequency of their feedback, the researcher met with them prior to the study to explain its objectives and discuss the feedback process in a training session. During training, teachers received a feedback booklet based on *Criterion* feedback, organized by the researcher, to guide their feedback provision. They were also provided with writing samples to analyze and offer feedback on, which they subsequently compared with the automated feedback from *Criterion*. Through practicing on samples and comparing their feedback with *Criterion*'s, the teachers were able to align the quantity and content of their feedback with that provided automatically by *Criterion*. The teachers also agreed to adhere to uniform practices to ensure consistency.

The remaining three instructors were tasked with scoring the pretest, immediate posttest, and delayed posttest. To ensure consistency, the grading process was standardized by providing the instructors with an analytic rubric (see Appendix A) to guide their scoring. Each teacher also

received examples of corrected samples provided by the researcher, to maintain consistency in the grading procedure.

Additionally, all participating teachers attended a two-hour teleconference training workshop led by an ETS-certified trainer before the study. This training provided them with detailed instructions on using *Criterion* as an automated writing evaluation system. In the workshop, the teachers were provided with instructions on how to register as a new user, add a class, create an assignment, add students to a class, work with students' portfolios, view reports on students' performance (see Appendix B).

4.5 Data Collection

4.5.1 Instruments

To obtain the data that provides answers to the research questions presented earlier, writing tests were used as an instrument. Being quasi-experimental, this study used pretest, immediate posttest, and delayed posttest to investigate any observed changes in participants' writing after the treatment sessions.

4.5.1.1 Writing Tests. Writing tests are widely regarded as reliable instruments for gathering written language samples within controlled time constraints. They are considered one of the common instruments used to gather written language samples under a specific time constraint (Hyland, 2016). In addition, they allow researchers to collect consistent and comparable data from participants in a systematic manner. Hyland (2019) stated that information obtained from written texts allows comparison between groups and/or previous learning performance. In the context of this study, the use of writing tests was specifically chosen to explore the impact of providing different modes of feedback on EFL students' writing outcomes.

In each of the three tests, participants were presented with two different prompts, each corresponding to one of two genres: expository and persuasive. In expository writing, students provide comprehensive information on a given topic, including descriptions, examples, and reasons, without favouring any particular point of view. In persuasive writing, students articulate their viewpoint on a specific subject, supported by reasons and examples. These genres were selected based on their relevance to the participants' proficiency level and the curriculum they are studying, as well as their alignment with the genres commonly encountered in the IELTS exam. The topics of the tests were devised by the researcher, drawing inspiration from the themes and topics covered in the students' curriculum book, "Life; Intermediate; National Geographic Learning."

The length and the time of the written text were controlled as well. The time constraint and word limit were imposed by the time and word limit given for the examinees of the IELTS exam. In the IELTS writing section, the examinees are given 60 minutes to produce two written texts which were similar to what was used in the current study. The IELTS examinees were advised to spend 20 minutes on the first writing task (150 words) and 40 minutes on the second writing task (250 words). The participants in the current study were then required to write approximately 150 words as an expository paragraph in 20 minutes and approximately 250 words as a persuasive essay in 40 minutes under exam conditions (see Appendix C). The comparison to the IELTS test serves a purpose, as it is the most commonly used assessment for testing the English language proficiency of foreign language users. Additionally, it is accredited for University of Jeddah students who seek certification that validates the English language course they must complete during their foundation year.

Test prompts used at the pretest, immediate posttest, and delayed posttest are as follows:

Version A:

Task 1: Reasons for Attending College (Expository)

People attend a college or university for many different reasons (for example, new experiences, career preparation and increased knowledge). Why do you think people attend college or university? Use specific reasons and examples to support your answer.

Task 2: Teenage or Adult life (Persuasive)

Some people think that the teenage years are the happiest times of most people's lives. Others think that adult life brings more happiness, in spite of greater responsibilities.

Version B:

Task 1: Why Study Abroad? (Expository)

Many students choose to attend schools or universities outside their home countries. Why do some students study abroad? Use specific reasons and details to explain your answer.

Task 2: Experience or Books (Persuasive)

It has been said, "Not everything that is learned is contained in books." Compare and contrast knowledge gained from experience with knowledge gained from books. In your opinion, which source is more important? Why?

Version C:

Task 1: Fictional Character (Expository)

Fictional characters from any genre (whether in books, movies, video games, etc.) often prove to be unforgettable. Write an essay about any fictional character that has had an effect on you. Fully describe the character, where you discovered him or her, and the effect he or she has had on you.

Task 2: Learning A New Language (Persuasive)

People who are learning a foreign language can face a number of difficulties.

What are some of these problems? In your opinion, what are the best ways to overcome these difficulties?

To further ensure fairness and consistency, Latin square counterbalancing of the topics was followed across the three testing phases (pretest, immediate posttest, and delayed posttest). This was necessary to avoid the effect of training and to give control over topic difficulty. A record of examinees was maintained to guarantee that each student responded to different prompts at each phase. For instance, a student in Group 1, Sub-group 1, completed **Test A** during the pretest, **Test B** at the immediate posttest, and **Test C** at the delayed posttest (see Table 2). This approach allowed for comprehensive evaluation while preventing repetition or familiarity effects across the testing phases.

Table 2 *Tests Counterbalancing.*

Test	Group		
	TF Group	AF Group	**A**TF Group
Pre-test	Sub-group 1 Test A	Sub-group 1 Test A	Sub-group 1 Test A
	Sub-group 2 Test B	Sub-group 2 Test B	Sub-group 2 Test B
	Sub-group 3 Test C	Sub-group 3 Test C	Sub-group 3 Test C
Immediate Post-test	Sub-group 1 Test B	Sub-group 1 Test B	Sub-group 1 Test B
	Sub-group 2 Test C	Sub-group 2 Test C	Sub-group 2 Test C
	Sub-group 3 Test A	Sub-group 3 Test A	Sub-group 3 Test A
Delayed test	Sub-group 1 Test C	Sub-group 1 Test C	Sub-group 1 Test C
	Sub-group 2 Test A	Sub-group 2 Test A	Sub-group 2 Test A
	Sub-group 3 Test B	Sub-group 3 Test B	Sub-group 3 Test B

The three tests were paper-based and were administered under exam conditions. To ensure reliability, three different markers who had the same qualifications marked the three tests following the same analytic rubric (see Appendix A) adapted from Connor-Linton & Polio, 2014). In this rubric, a total mark out of 100 for each task (i.e., task1 for expository genre (100), task2 for persuasive genre(100)). The total mark for each task was divided as well between five components: text content(20), text organization(20), vocabulary use(20), language use,(20) and mechanics(20). The total mark of each test was 200.

4.5.2 Scoring

All scoring was conducted blindly, and all participants' information was kept hidden to ensure confidentiality. Three English language teachers marked students' papers according to the rubric provided (see sub-section 4.6.2.3). After grading, the three scores assigned by the three different markers were compared. If a significant discrepancy existed between the markers, the researcher re-evaluated the same paper. Otherwise, the average of the three scores was considered the final mark for the exam.

Upon review, it was found that the grades given were higher than anticipated. Consequently, the students' papers were re-evaluated by me. Another PhD researcher from the same field marked 20% of the papers. Subsequently, an intraclass correlation coefficient test (ICC) using the One-way model was conducted to assess the interrater reliability of the scores. The *ICC* assesses the proportion of total variance that can be attributed to the differences between the raters compared to the total variance. It is suitable for assessing agreement between two or more raters when the data is continuous. The rationale for checking inter-rater agreement was to ascertain that the two raters can reach an acceptable level of agreement using the same rubric to judge the quality of participants' written texts and thus achieving an acceptable level of evaluation reliability. The result was found to be 0.98, which suggests an excellent level of agreement between the two raters in their measurements (see Table 3). The 95% confidence interval for the *ICC* population values was calculated to be between 0.98 and 0.99, further supporting the robustness and precision of the estimated *ICC* value. In conclusion, the single score intraclass correlation analysis demonstrated highly reliable and consistent measurements between the two raters with an *ICC* value of 0.98, thereby indicating excellent inter-rater agreement, (Landis and Koch, 1977).

Table 3 *Single Score Intraclass Correlation Coefficient (ICC).*

Model	Type	Subjects	Raters	*ICC*(1)	*F*-Test	*F(df1, df2)*	*p*-value	95% *CI* for *ICC*
Oneway	Agreement	540	2	0.984	H0: r0 = 0	F(539, 540)	0	0.981 < *ICC* < 0.986

4.5.3 Treatment Sessions

During the treatment sessions, participants produced four written texts across two genres: two expository essays and two persuasive essays. These tasks were adapted from curriculum materials and carefully selected to correspond with the participants' proficiency levels. Engaging in written production is widely recognized as an effective approach for enabling students to revise their work based on the feedback they receive (Hyland, 2019; Hyland & Hyland, 2006b, 2006a)[1].

Following are the prompts which the participants responded to during the treatment sessions:

[1] For more details on the writing tasks and the feedback sessions procedure, see the following section 4.5.3.

Session 1: Good Friend (Expository)

What are the qualities of a good friend? Write an essay in which you describe what it takes to be a good friend. Identify the qualities a person must have to be a good friend, and develop those ideas with specific examples and support, citing your own experiences.

Session 2: Money on Technology (Persuasive)

Some people think that governments should spend as much money as possible on developing or buying computer technology. Other people disagree and think that this money should be spent on more basic needs. Which one of these opinions do you agree with? Use specific reasons and details to support your answer.

Session 3: New Product (Expository)

If you could invent something new, what product would you develop? Use specific details to explain why this invention is needed.

Session 4: Change Job or Not (Persuasive)

Some people prefer to change jobs or professions during their careers. Others choose to stay in the same job or profession. Discuss the advantages of each choice. Which do you prefer? Use reasons and examples to explain your choice.

4.5.4 Procedure

A detailed description of the procedure is explained in the following section under the subheadings pre-treatment stage, treatment stage, and post-treatment stage.

4.5.4.1 Pre-treatment Stage. Week 1 was dedicated to preparing the context for the study. First, all the student participants completed an Oxford Placement test (http:// www. oxfordenglishtesting.com/). They were then given a brief introduction to the purpose of the study, their role as participants, and the general instructions provided by the researchers.

Afterward, the participants signed their consent forms. Finally, the participants completed the writing pretest under exam condition.

The participants in each class were randomly divided into three groups: automated feedback, teacher feedback, and hybrid feedback. This division was based on the type of feedback they would receive during the treatment (see Figure 2). After the groups were established, the participants were provided with instructions on how the treatment sessions would be conducted. The groups using the automated evaluation system, *Criterion*, had an additional session where they were introduced to the system and its functionalities. Participants received a *Criterion Student Access Guide*, which contains detailed information on how to register and log in as a student, begin their responses, view feedback from *Criterion*, revise their submissions, access teacher and peer feedback (if applicable), utilize available help and resources, and archive their portfolios. After this, participants were provided with the necessary usernames and passwords to log into the system. They were then given time to practice using *Criterion* to ensure they could navigate it confidently and avoid any technical issues during the treatment sessions.

4.5.4.2 Treatment Stage. The treatment sessions took place once a week for four successive weeks (weeks 2,3,4,5). Each week they had one writing session divided into two parts with a 30-minute break in between. In the first part of the session (30 minutes), the participants wrote a first draft about the topic provided. They had a 30-minute break then to allow the assigned teacher to provide feedback on writing to the participants in TF group (less than 10 students). After that, all participants spent the second part of the session (50 mins) revising and reproducing final drafts after receiving their feedback. The TF group received feedback directly from the teacher, who provided written feedback on their papers. The AF group received

automated feedback via *Criterion* while revising their work. The ATF group received teacher feedback only on content and organization, inserted as comments through *Criterio*n, while simultaneously using automated feedback to revise other components of their writing. The same procedure repeated through the four writing sessions over four weeks. This was done to ensure that all participants had an equal amount of time dedicated to revising and rewriting their texts. The specific procedure for each group during the treatment sessions was as follows:

Teacher Feedback Group. This group consisting of 25 students, received written feedback from their teacher on their writing. They used pen and paper to write the texts and received teacher written feedback directly on their papers. This group also served as a control group, as it adhered to the conventional feedback condition (teacher feedback) typically employed in writing instruction at the University of Jeddah. Assigning a control group is crucial for gauging the impact of extraneous variables and ensuring that any progress observed in the intervention is attributable to the independent variable (Kumar, 2019). The teacher was provided with a rubric (see Appendix B) outlining the writing aspects to be emphasized during the feedback session (i.e., text content, text organisation, vocabulary use, language use, and mechanics). The rubric was adopted from the *Criterion* system which provided automated feedback to the other participants. This measure was taken to ensure consistency in feedback practices across all groups during the treatment sessions (refer to the teacher's training section for further details on the steps taken by the researcher to ensure consistency).

Automated Feedback Group. This group, also had 25 students, exclusively received automated feedback provided by *Criterion*. *Criterion* is automated writing evaluation system used in the current study. Using iPads and the university's Internet access, the students logged into *Criterion* system to practice writing and receive automated feedback. The prompts were

delivered electronically through the platform. The teachers' responsibilities for this group included initiating discussions on the topic during the first session, reporting any technical issues or difficulties related to the use of *Criterion*, and ensuring that students submitted their writing productions within the allocated time frame.

Hybrid Feedback Group. This group had 24 students. They received automated feedback provided by *Criterion* on different linguistic features (grammar, mechanics, usage, and vocabulary), while feedback on text content and organization was deliberately disabled in the system (*Criterion* allows the teacher to select the categories on which the system should provide feedback). Feedback on text content and organization then became the teacher's responsibility (see Figures 3 and 4). Consequently, the teacher's role in this group involved initiating discussions on the topic during the first session, reporting any technical issues or difficulties related to the use of *Criterion*, providing feedback on text organization and content through the system during feedback and revision sessions, and ensuring that students submitted their writing productions within the allocated time.

The decision to allocate content and organization feedback to teachers was intentional. As discussed in the literature reviewed, current AWE systems such as Criterion are limited in their ability to evaluate idea development, coherence, and rhetorical structure. Research (e.g., Weigle, 2013; Wilson & Czik, 2016) has consistently shown that while AWE can reliably flag linguistic errors, it lacks the sophistication to provide meaningful feedback on higher-order writing concerns. For this reason, teacher feedback was prioritized in these areas, since human judgment is still necessary to guide learners in developing arguments, supporting ideas, and improving essay structure.

To ensure that the three groups receive an equal amount of feedback, the time for each

session during the intervention was fixed for all participants. All groups used the same prompts provided by the researcher during the treatment sessions. Teacher feedback was written on teacher feedback group papers during the 30-minute break between the writing and revision sessions. The participating teachers had a training session with the researcher to ensure they followed the same practice during the feedback sessions (see 4.4.2).

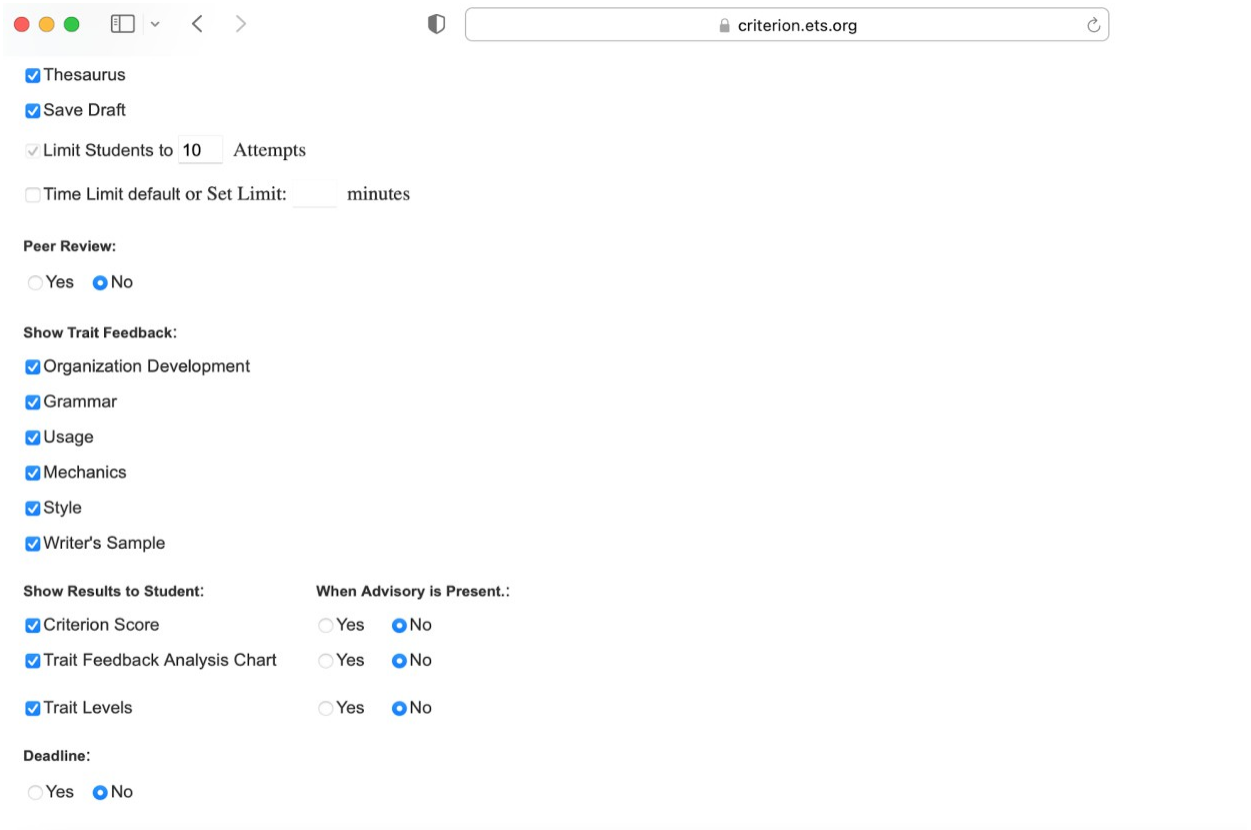


Figure 3 Trait Feedback in Criterion.

Note: *Trait Feedback selection was used to disable automated feedback on text organisation and content.*

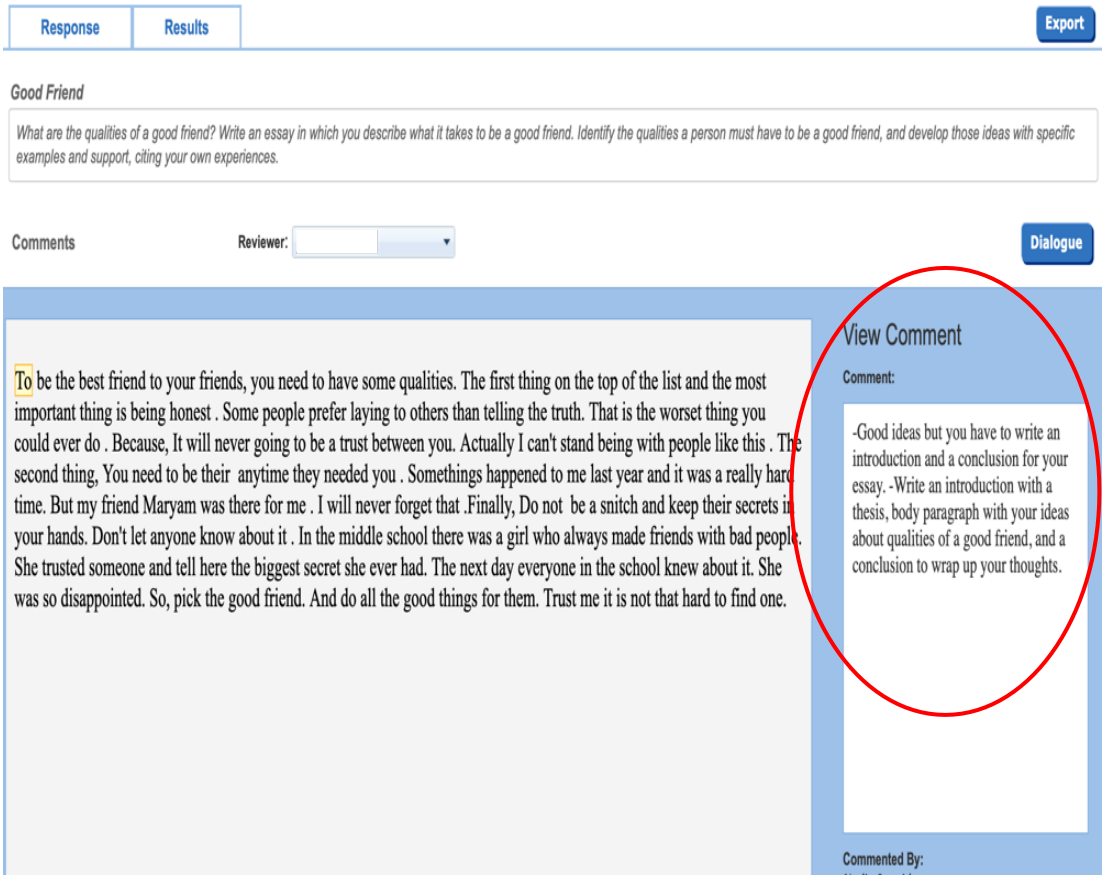


Figure 3 Example of Teacher Feedback Inserted in Criterion for the Participants (Hybrid Feedback Group).

4.5.4.3 Post-treatment Stage. After the treatment sessions, all participants had an immediate posttest in writing a day after the last treatment session (Week 5). Finally, they completed a delayed posttest (on week 10, five weeks after the intervention) to ensure they had retained the treatment they received. Figure 5 diagrammatically presents the design of the study. The three writing tests (pretest, immediate posttest and delayed posttest) had the same prompts. However, as mentioned earlier Latin square counterbalancing of the topics was followed to avoid the effect of training and to give control over topic difficulty (see Table 2).

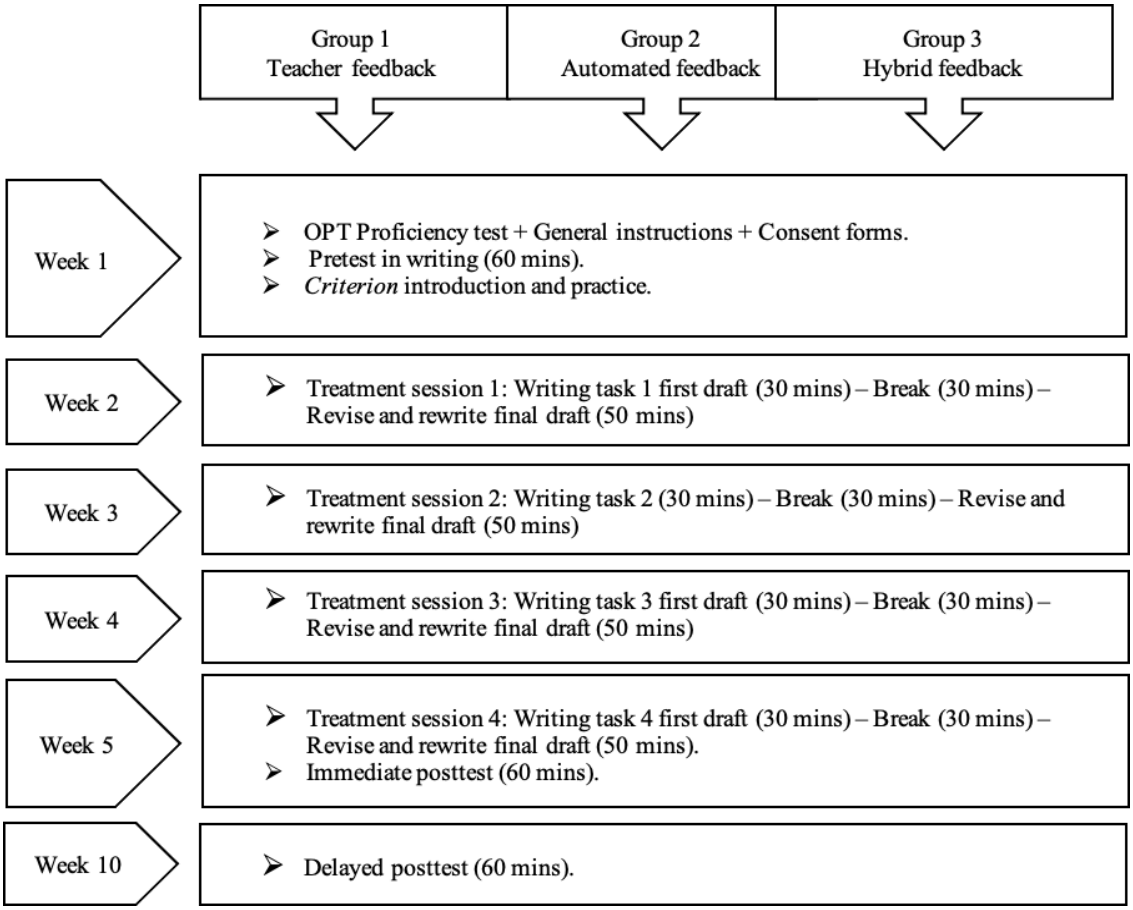


Figure 4 A diagram of the Study Design.

4.5.5 Criterion Software

Criterion is the automated writing evaluation system used by two groups of participants during the intervention. It highlights errors and provides feedback according to different categories (see Appendix B). It also offers comments on text content and organization (Attali, 2004; Attali & Burstein, 2006). Furthermore, *Criterion* provides a wide range of individualized feedback (Warschauer & Ware, 2006), this feature is thought to be beneficial as it gives individualized attention to each student (Hyland & Hyland, 2006b). In addition, *Criterion* software uses the e-rater scoring system which is currently used in more than 20 applications besides *Criterion* including: GRE, TOEFL, TOEFL online practice (Lim & Kahng 2012).

It also relates to a large corpus of edited texts to detect any "violation" of general English including (Attali, 2004). In addition, it provides feedback about discourse elements like the introduction, thesis statement, main and supporting ideas, and conclusion as it is also linked to a large corpus of human annotated essays (Burstein et al., 2003). Warschauer and Ware (2006) examined the features of different AWE programs and found out that *Criterion* can provide feedback to prompts which have not been normed by human scores. This feature allows teachers to add different prompts to correspond to the curriculum they are teaching. In addition, the studies that examined *Criterion* scoring system concluded that there was a sound agreement in the scores generated by *Criterion* and human raters (Attali, 2004; Condon, 2013; Powers et al., 2015; Shermis et al., 2008). For example, Power et.al. (2015) found that the scores awarded by *Criterion* agree more with trained raters rather than untrained raters and that increased the reliability of the system. The use of *Criterion* platform was also favored by students and motivated them to do more revisions (Shermis et al., 2008). Moreover, *Criterion* is extensively investigated in the literature.

A recent systematic review by Loncar et al. (2023) found that 46% of the studies examined used *Criterion* to assess the effectiveness of automated writing evaluation.

Additionally, *Criterion* platform offers flexibility to the instructor in terms of usage and control over students' writing. It gives the instructor the ability to control the type of feedback provided to students. For example, teachers can disable some features and enable the software to focus on desired feedback traits (see figure 3). This is an important feature for the current study to distinguish between the type of feedback provided for each group.

4.6 Data Analysis

This section presents the data analysis procedures undertaken to analyze the data obtained through the writing tests (pretest, immediate posttest, and delayed posttest). The analysis followed a quantitative research method which allows the investigation of a large number of cases in an appropriate timeframe (Flick, 2020). This investigation was set to measure the changes in students' writing quality under three feedback conditions (TF, AF, ATF). In addition, the study explored whether change in writing under any of the three feedback conditions was affected by the change in writing genre (expository, persuasive). Writing genre was treated as a moderator variable, to examine if it may influence the strength or direction of the effect of feedback on writing outcomes.

The quality of writing was measured using an analytic rubric (Connor-Linton & Polio, 2014). The independent variable in this study was the three conditions of feedback on the students' essay writing plus the writing genre type. The dependent variable was the change in students' writing (see Table 1). The dependent variable was an interval variable, discrete or continuous, calculated based on the scores awarded to a pretest, immediate posttest, and delayed posttest.

Written tests were scored according to different components of a written text: text content, text organization, vocabulary use, language use, and mechanics using a revised analytic rubric (Connor-Linton & Polio, 2014). Each of these aspects was weighted out of 20, giving a total of 100 for each writing task. Each test consisted of two writing tasks (expository and persuasive), making the total score of each test 200. Inter-rater agreement was checked and discussed in the previous section (see Table 3).

Data were compiled into an Excel spreadsheet with columns arranged using the long format to be uploaded into R software for statistical analysis (R Core Team, 2022). Descriptive statistics were initially used to summarize and visualize the data with the *tidyverse* package (Wickham, 2016). Before selecting the appropriate inferential test, the normality of the data was assessed using the Shapiro-Wilk test. Results indicated that not all data followed a normal distribution, a finding further supported by QQ plots and histograms (see Appendix E, F).

As a result, mixed-effects models were selected for data analysis due to their flexibility and robustness compared to traditional analysis of variance (Cunnings & Finlayson, 2015). Specifically, linear mixed-effects models exhibit robustness against normality assumptions. As indicated by its name, mixed-effects model enables combining 'fixed effects' and 'random effects'. Fixed effects can be continuous or categorical variables, whereas random effects are necessarily categorical (Winter, 2020). The incorporation of random effects such as test version, instructors, and test timing permit the examination of any potential impact of these factors on the outcome.

4.6.1 Operationalisation of Measures

In this section, the chosen dependent variables for assessing the impact of the treatment variables are explained (see Table 4). Dependent variables were explicitly operationalized here to evaluate changes in the students' writing.

Table 4 *Explanation of Variables.*

Independent variable	Dependent variable
Categorical data: 1. Feedback on writing a categorical variable with three levels: automated feedback; teacher feedback; teacher and automated feedback. 2. Test time categorical variable with three levels: pretest, posttest, delayed posttest. 3. Writing genres categorical variable with two levels: expository; persuasive.	Numeric data: 1. Overall writing score (/200) on pretest, immediate posttest and delayed posttest. 2. Scores for components of writing: text content (/40), text organisation (/40), vocabulary use (/40), language use (/40), and mechanics (/40). 3. Total score calculated separately for each genre: expository (/100), persuasive (/100).

The dependent variable, total writing score (weighted at 200 points: 100 for expository writing and 100 for persuasive writing), was measured as the sum of evaluations across five writing aspects: text content, text organization, vocabulary use, language use, and mechanics. Each aspect was independently scored out of 20 points, with all aspects weighted equally for each task (expository and persuasive writing). The total score for each writing aspect was 40 points (20 for expository and 20 for persuasive writing).

This evaluation process was conducted at pretest, immediate posttest, and delayed posttest phases for the three feedback groups. These scores were subsequently analyzed both within groups and between groups.

In assessing students' papers, the markers followed an analytic rubric (see Appendix A). This rubric has been adopted in numerous previous studies (e.g., Connor-Linton & Poilo, 2014; Martínez, 2018; Mazgutova & Kormos, 2015). The criteria for each aspect are explained below:

1. Text content: According to the rubric, the content of a well-written essay should feature the following: 1) a thorough and logical development of the thesis; 2) a substantive and detailed body of work; 3) no irrelevant information, 4) interesting content, 5) a substantial number of words for the time given.
2. Text organisation: This refers to how a text is organised and developed logically. In other words, the steps the writer implements to make the text understandable to the readers. According to the current rubric, a well-organised text has: 1) excellent overall organisation; 2) a clear thesis statement; 3) a substantive introduction and conclusion; 4) excellent use of transition words; 5) excellent connections between paragraphs; 6) unity within every paragraph.
3. Vocabulary use: Similar to text content and text organisation, the rubric marks the appropriate use of vocabulary by: 1) the use of sophisticated vocabulary; 2) the choice of words with no errors; 3) an excellent range of vocabulary; 4) idiomatic and near native-like vocabulary; 5) an academic register.
4. Language Use: This section refers to the correct use of language with regards to sentence structure. The rubric also defines good use of language which includes:1) no major errors in word order or complex structure; 2) no errors that interfere with comprehension; 3) only

occasional errors in morphology; 4) frequent use of complex sentences; 5) excellent sentence variety.

5. Mechanics: This function is also operationalised through numerous aspects: 1) an appropriate layout with indented paragraphs; 2) no spelling errors; 3) no punctuation errors.

4.6.2 One-way ANOVA

To address the proposed research questions, R software (R Core Team, 2022) was employed to conduct two types of statistical analysis: descriptive and inferential. As a preliminary step, an analysis of variance (ANOVA) was performed to ensure the comparability of groups prior to the experiment. ANOVA was used to identify any significant differences between groups in the total scores of the pretest and the Oxford Placement Test (OPT). Additionally, it assessed the comparability of groups in each of the evaluated writing components—text content, text organization, vocabulary use, language use, and mechanics—before the intervention.

4.6.3 Descriptive Statistics

Descriptive statistics were generated and plotted using the *tidyverse* package (Wickham, 2016). The means, standard deviations, medians, and interquartile ranges were reported to provide an overview of the overall students' writing performance at the three tests time (pretest, immediate posttest, delayed posttest) relative to the three feedback conditions.

Descriptive statistics were also used to provide an overview of the data in relation to each of the components of writing examined (i.e., Text content, text organization, vocabulary use, language use, and mechanics). Additionally, descriptive statistics were reported separately for each feedback group based on the writing genres assessed (expository and persuasive).

4.6.4 Inferential Statistics

Data were analyzed using linear mixed-effects models with the *lme4* package (Bates et al., 2015). This statistical method was chosen because the dependent variable, scores, was continuous (Winter, 2020). The dependent variable (Score) refers to the overall writing test score (/200), and was examined in the first research question.

The model structure included feedback condition (teacher feedback, automated feedback, hybrid feedback) and test time (pretest, immediate posttest, delayed posttest) as fixed effects, with an interaction between them.

Feedback conditions were dummy coded. Teacher feedback group was dummy coded as 0, automated feedback group as 1, and hybrid feedback group was coded as 2. Participants, teachers, and test versions were set as random effects. (Subject) which refers to the students' participants was set as random intercept and random slope for test time. Teacher was set as random intercept and random slope for the interaction between feedback type and test time. Test version was set as a random slope for the interaction between feedback type and test time. The structure of the syntax using *lme4* was set as follows:

```
Score~Test*Group + (1+Test|Subject) + (1+Test*Group|Teacher) + (Test*Group|Version)
```

Buildmer package was used then to find the maximal feasible model which would include all possible random effects that may affect the model's outcome (Barr et al., 2013). The *LmerTest* package (Kuznetsova et al., 2017) was used to calculate *p*-values in the linear mixed-effects models. Afterwards, Cohen's *d* effect sizes and 95% confidence intervals were reported to provide deeper understanding of the practical implications of the findings.

For the second research question, the dependent variable analyzed was the components of writing. To investigate these aspects, the linear mixed-effects model was run five times, each time

focusing on a different writing component: 1. text content (/40), 2. text organization (/40), 3. vocabulary use (/40), 4. language use (/40), 5. mechanics (/40). The structure of the syntax using *lme4* was set as follows:

1- ContentScore~Test*Group + (1+Test|Subject) + (1+Test*Group|Teacher) + (Test*Group|Version)
2- OrganizationScore~Test*Group + (1+Test|Subject) + (1+Test*Group|Teacher) + (Test*Group|Version)
3- VocabularyScore~Test*Group + (1+Test|Subject) + (1+Test*Group|Teacher) + (Test*Group|Version)
4- LanguageScore~Test*Group + (1+Test|Subject) + (1+Test*Group|Teacher) + (Test*Group|Version)
5- MechanicsScore~Test*Group + (1+Test|Subject) + (1+Test*Group|Teacher) + (Test*Group|Version)

For the third research question, three linear mixed-effects models were run using the *lmer* in the *lme4* package for each feedback condition. Total writing score was the dependent variable. Writing genre (expository, persuasive) was added to test time (pretest, immediate posttest, delayed posttest) with an interaction between them as fixed effects. Subject was set as random intercept and random slope for writing genre. Teacher was set as random intercept and random slope for the interaction between writing genre and test time. Test version was set as a random slope for the interaction between writing genre and test time. *Buildmer* package was used then to find the maximal feasible model which would include all possible random effects that may affect the model's outcome (Barr et al., 2013). The *LmerTest* package (Kuznetsova et al., 2017) was used to

calculate *p*-values. Linear regression outcomes were then plotted using *alleffects* function from the *effects* package (Fox and Weisberg, 2018) to visualise and compare the results. The syntax structure for the *lme4* models was as follows:

1- TF_Score~Test*Genre + (1+Test*Genre|Teacher)+ (Test*Genre|Version) + (1+Genre|Subject)

2- AF_Score~Test*Genre + (1+Test*Genre|Teacher)+ (Test*Genre|Version) + (1+Genre|Subject)

3- ATF_Score~Test*Genre + (1+Test*Genre|Teacher)+ (Test*Genre|Version) + (1+Genre|Subject)

4.6.5 Reporting Results

For each research question, the analysis began with an ANOVA to confirm group comparability prior to the treatment. Next, descriptive statistics were reported to provide an overview of participants' performance across conditions. This was followed by inferential statistics to examine whether the treatment had a statistically significant effect on EFL learners' writing performance.

Inferential statistics were calculated using mixed-effect models in R. Mixed-effects models are used over ANOVA for several reasons. Mixed-effects models were chosen for data analysis because they offer greater adaptability and statistical reliability than traditional ANOVA techniques (Cunnings & Finlayson, 2015). Unlike ANOVA, linear mixed-effects models are less sensitive to violations of normality assumptions, making them more robust in handling data. These models also provide the advantage of simultaneously accounting for both fixed effects such and random effects (Winter, 2020). By incorporating random factors like test version, instructor, and testing time, mixed-effects models allow for a more nuanced analysis of how these variables may influence the dependent outcome, something that traditional ANOVA does not accommodate as effectively.

It is worth mentioning that there are two types of mixed logistic regression: generalised linear mixed models (GLMMs) and linear mixed models (LMM). The former is suitable for modelling binary responses while the latter is applicable to continuous responses. As responses in the present study are continues, LMM is the model of choice. The function that performs LMM in lme4 package is *lmer*. After applying *lmer*, *Buildmer* package was used then to find the maximal feasible model which would include all possible random effects that may affect the model's outcome (Barr et al., 2013). The *LmerTest* package (Kuznetsova et al., 2017) was used to calculate *p*-values. The results are retrieved by requesting a summary of the model. The summary of model includes values for (estimate, standard error, degree of freedom, t value, and p value). Following is a commentary adapted from Tagliamonte (2012) and Winter (2020) on how to interpret mixed-effects model summary outcome values (see table):

1- Fixed factor/ effect: Variables that are of primary interest and whose effects are estimated across the entire population. Can be continuous or categorical. In the current study fixed effects were feedback groups, test time, and genre.

2- Random factor/ effect: Variables that help control for this nested or repeated structure in the data, allowing for more accurate and generalizable estimates of the fixed effects. Always categorical and not the main focus of the analysis. For example, instructor, test version, and subject (participants) were the random effect in the current investigation. They were not the main focus but they might influence the main variables.

3- Intercept: baseline value of the dependent variable before the effects of the predictors are added. In linear mixed effect model, the intercept is the average score of the reference group which is compared then to the other levels. In table (00), the intercept

is set as the teacher feedback (TF) group at the pretest; however, it can be redefined as needed to enable different comparisons.

4- Estimate: The predicted effect of a variable on the outcome. For example: the estimate of 11.16 means the average score of the teacher feedback group (TF) at the pretest. The estimate of 7.36 means that the outcome of TF group increased by 7.36 units at the immediate posttest (11.16 + 7.36) and by 6.26 units at the delayed posttest (11.16 + 6.26). If an estimate is negative (e.g., -6.26), it indicates a decrease from the baseline (11.16 – 6.26). The outcomes of AF group and ATF group are interpreted relative to the intercept, which corresponds to the TF group at the pretest.

5- Standard error (SE): The standard deviation of the estimate, indicating how precise the estimate is. A smaller SE means the estimate is more precise.

6- Degrees of freedom (df): The number of independent pieces of information used to calculate the estimate.

7- t-value: Reflects the number of independent pieces of information used to calculate the estimate. Higher absolute t-values suggest a stronger effect.

8- p-value: Shows whether the effect is statistically significant (commonly if $p < .05$). Lower p-values indicate the result is less likely due to chance.

Table 5 *Summary of Fixed Effects Estimates (Reference: Teacher Feedback Group and Pretest): Example for Illustration.*

Fixed effects	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	11.16	1.71	9.938	6.52	6.94E-05***
Posttest	7.36	1.47	101.447	4.99	2.45E-06***
Delayed posttest	6.26	1.93	111.483	3.24	0.00158**
AF group	2.97	2.07	126.818	1.44	0.15361
ATF group	1.92	2.08	126.26	0.92	0.35725

In addition to *p*-values, Cohen's *d* effect sizes were reported to indicate the magnitude of observed effects. The significance level (alpha) in the current study was adjusted for multiple comparisons across repeated tests to control for Type I and Type II errors. For example, in reporting the mixed-effects model results, *p* value was adjusted to account for four repeated tests in the first and second research question (*p<.0125. **p<.0025. ***p<.00025) and two repeated tests in the third research question (*p<.025. **p<.005. ***p<.0005). Furthermore, statistical significance was interpreted alongside other measures, such as confidence intervals and effect sizes, to provide a more comprehensive understanding of the results.

Reporting effect sizes alongside *p*-values offers several advantages. Effect size metrics such as Pearson's *r* or Cohen's *d* reflect variability in the sample (Winter, 2020). Moreover, effect sizes are not influenced by sample size; as such, meaningful effects can still be detected

even in smaller samples. Therefore, effect sizes complement *p*-values by indicating the practical significance of the results.

In this study, Cohen's *d* was used to estimate effect sizes. The interpretation of these values followed the L2-specific thresholds proposed by Plonsky and Oswald (2014), which are considered more appropriate for the current study than the general benchmarks suggested by Cohen (1988). For within-subject differences, small, medium, and large effects correspond to values of .60, 1.00, and 1.40, respectively. For between-group differences, the thresholds were .40 (small), .70 (medium), and 1.00 (large).

4.7 Pilot Study

The purpose of the pilot study was to test the instruments before employing them in the main study. It also assisted in identifying any problems that might hinder participants' understanding of the tests and the procedures of the interventions (Kumar, 2019). Furthermore, it addressed any issues and allowed the researcher to make adjustments before conducting the main study. Additionally, it provided some preliminary statistics to indicate whether the effects are likely to be present.

The actual pilot study was conducted approximately two weeks before the main data collection process at the same research site. Initially, the intention was to recruit 30 female students to participate in the pilot study, but this could not be achieved due to reasons pertinent to the study context, as they were undergoing mid-term exams. In addition, University of Jeddah was following post-pandemic procedures at the time of the study and most of the students were attending virtually which added to the difficulty to assign large number of participants. Consequently, only six participants agreed to participate at that time, and they were randomly assigned to three groups following the same procedure intended for the main data collection process. No data analysis was performed for the pilot study due to the small number of

participants. However, piloting the data made it possible to test the instruments and ensure that the participants understood the procedure. This process provided an opportunity to make certain changes regarding the plan for data collection.

4.8 Validity and Reliability

Validity and reliability are vital to any research design (Cohen et al., 2018; Creswell & Gutterman, 2019; Shadish et al., 2002; Winter, 2000). Validity concerns with truthfulness in terms of investigating exactly what is claimed to be investigated. Reliability, on the other hand, is concerned with the replicability of the data (Winter, 2000). It is very important to obtain valid and reliable results from research. Creating valid research means that the researcher has the confidence to conclude that the treatment itself is responsible for the observed effects and not chance or some other confounding factor, such as practice, maturation, or measurement problems. To have a reliable study means the study can be confirmed and reproduced in other context or in other time by the researcher himself or someone else. Validity and reliability increase the quality of research and can be achieved through careful attention to how the study is designed.

Internal validity, on the other hand, refers to the degree of confidence that the causal relationship being tested is not influenced by other factors or variables. (Cohen et al., 2018). In other words, the findings should be directly related to the variables set by the researcher. In experimental studies, for example, a causal relationship should be drawn between the treatment variables and the outcome (Shadish et al., 2002). Issues related to internal validity could compromise this relationship and resulting in the experiment failure. To illustrate, the findings of the experiment may not relate to the variables set by the researcher and thus will not answer the research questions properly. According to Cohen et al. (2018) history, maturation, ambiguity in

the process, statistical regression, testing, instrumentation, and selection are some of the threats that might affect internal validity in quantitative research especially in experiments.

Threats to validity are possible and may affect any research method, whether quantitative or qualitative. Therefore, several procedures were followed to mitigate those threats in the current study. First, a pilot study was conducted prior to the main data collection to ensure that the instruments designed to address the research questions accurately measured the relationship between the variables and allowed for modifications to the design's errors prior to the experiment.

The present study adopted quasi-experimental research deign where internal validity is inferior to that of true experiments due to lack of randomisation of sample (Dörnyei, 2007). Notwithstanding, the choice of convenient sampling in the current study (three intact classes) provides ecological validity to the research as it reflects the natural context of participant and their daily experience (Gass & Mackey, 2015)

Furthermore, criterion validity was attended to through type of data collected. Data were collected through writing tests which is considered as a valid source of data and was used repeatedly in numerous previous studies to investigate the improvement of L2 writing. In addition, a pretest, immediate posttest, and delayed posttest design was adopted as the study sought to compare different feedback conditions. Moreover, validity of results was maintained through the use of a revised analytical rubric (Connor-Linton & Polio, 2014) which is theoretically based.

Reliability was also considered in the careful design of the study, referring to the quality of procedures that provide reproducibility and accuracy. Several steps were taken to enhance the reliability of the research. First, all participants completed the Oxford Placement Test (OPT) and

the writing pretest before the treatment sessions to ensure uniform proficiency levels. The study also controlled for other extraneous variables such as background (all participants were Saudi), age (participants aged 18-19 years old).

Second, participating instructors were trained and provided with a rubric to follow during the intervention and while scoring and providing feedback on writing. Additionally, three different instructors of similar qualifications were responsible for scoring the writing tests, and interrater reliability was considered and checked. All participants followed identical practices during treatment sessions and writing tests, as described in the procedure section.

Finally, several efforts were made in the current study to minimize bias. As noted earlier, each participating instructor was involved with the three different feedback groups. The researcher also observed the entire procedure to ensure uniformity in timing and duration of treatment stages for each group and to identify any potential sources of bias. Identical materials and tasks were used for all groups, and test counterbalancing was employed to mitigate the effect of repetition across the pretest, the immediate posttest, and the delayed posttest, which covered identical topics.

4.9 Ethical Considerations

Prior to the commencement of the pilot study, ethical approval was formally obtained from the University of York, ensuring that all aspects of the research adhered to the institution’s ethical guidelines. This approval process underscored the study's commitment to maintaining the highest ethical standards and protecting the rights and well-being of all participants.

Following the approval, a formal letter was sent to the Vice Dean and Head of the Research Unit at the English Language Institute, University of Jeddah. The letter outlined the study’s purpose, specific research objectives, and anticipated benefits, while also emphasizing

the researchers' adherence to ethical protocols and the importance of collaborative engagement. Permission was granted shortly thereafter, allowing the research to proceed with the active cooperation of the institution.

After receiving permission, a meeting was convened with the Vice Dean of the English Language Institute, the Head of the Research Unit, and the participant instructors at the University of Jeddah. This meeting served as a platform to discuss the study's overarching goals, methodological framework, and anticipated timeline in greater detail. During this session, the researchers also addressed questions and concerns from the attendees, ensuring that all parties had a clear understanding of their roles and responsibilities.

To further ensure a seamless and ethical implementation of the research, an orientation session was organized prior to the intervention. This session aimed to familiarize both the instructors and the participating students with the study's objectives, design, and specific procedures. Special attention was given to explaining the importance of informed consent, the voluntary nature of participation, and the measures in place to protect participants' privacy and confidentiality. During this session, the researchers emphasized the value of participant contributions to the study and provided clear instructions on how data collection and analysis would proceed.

Following this, the participants were invited to sign relevant consent forms. To ensure full comprehension, the students' consent forms were translated into Arabic to facilitate completion. Anonymity and confidentiality of information were reiterated and assured to all participant students and instructors. Codes were employed during data analysis instead of participants' real names. Finally, participants were informed of their right to withdraw from the study at any time, up to two weeks after data collection, without the need for explanation.

4.10 Summary

This chapter outlined the methodology and research methods employed in the present study. It began by presenting the research questions, followed by a detailed explanation of the study design. A comprehensive description of the study context and participants was then provided. Subsequently, the chapter detailed the data collection procedures. This was followed by a discussion of the data analysis methods and the operationalization of key variables, ensuring clarity in how the variables were measured and interpreted. The chapter also included an account of the pilot study, addressing its procedure and purpose, alongside a discussion of the study's validity and reliability measures. Lastly, it highlighted key ethical considerations, demonstrating the study's adherence to ethical research standards. Overall, this chapter aimed to provide a thorough account of the study design, data collection instruments, and procedures, establishing a foundation for the results and discussion presented in the subsequent chapter.

5. Results

5.1 Overview

This chapter presents the results of the current study by answering three research questions relating to the effect of feedback condition on student writing performance, and to the effect on different writing genres. The quantitative data were obtained from the participants' written texts at three test phases: a pretest, an immediate posttest and a delayed posttest. Due to the fact that the study compared the effects of three feedback conditions (teacher feedback only, automated feedback only and hybrid -automated+teacher- feedback) on participants' writing performance, it was necessary to address changes in overall writing production as well as changes in writing components (text content, text organisation, vocabulary use, language use and mechanics) over time. Additionally, the study examined changes in writing genres (expository and persuasive essay) and whether this affected the participants' writing performance relative to their feedback condition.

The chapter presents the results of the three research questions respectively. For each question, the results for the descriptive statistics (mean, standard deviation, median and interquartile range) are reported first to provide an overall summary of the results. Then the results obtained from the mixed-effects model are presented, along with the Cohen's *d* effect sizes and 95% confidence intervals to report between-group and within-group differences at the pretest, immediate posttest and delayed posttest.

5.2 Research Question 1: L2 Overall Writing Production Predicted Relative to the Feedback Condition

This section sets out to answer the first research question: To what extent do three different feedback conditions (teacher only, automated only, hybrid feedback) affect overall L2 writing production?

Initially, the analysis of variance (ANOVA) test was employed as the primary step before conducting data analysis to ensure the comparability of groups prior to the experiment. Descriptive statistics are reported next to provide an overview of the data regarding each group.

The data were analysed using linear mixed-effects models with feedback condition and test time (with an interaction between them) as fixed effects. The subjects (participants) were set as random intercept, teacher and test version were set as the random slope for the interaction between the feedback type and test time. The *Buildmer* package was used to find the maximal feasible model which would include all of the possible random effects that may affect the model's outcome (see Data Analysis Methodology section). The results for the models are presented, focusing on the model estimates (β) which directly address research question one, with a 95% confidence interval and Cohen's *d* effect sizes.

Regarding the overall writing scores, the ANOVA results suggested that there was no significant difference among the three groups at the pretest, F (2,71) = 1.34, p = .268. The results for the One-way ANOVA also indicated that there was no significant difference among participants on the Oxford placement test which was administered prior to the experiment, F (2,71) = 2.136, p =.126. Therefore, the overall writing and language competence was the same for all three groups and, thus, they were comparable and had similar proficiency at the start of the treatment.

Table 6 provides the descriptive statistics for the overall writing scores achieved by the participants in the three feedback condition groups at three test times. Descriptive statistics for the teacher feedback group revealed that there was a change in mean scores from the pretest (*M* = 47.7, *SD* = 34.3) to the immediate posttest (*M* = 85.4, *SD* = 35.2) and to the delayed posttest (*M* = 86.8, *SD* = 43.7). The automated feedback group also showed a change in mean scores from (*M* = 64.1, *SD* = 38.8) to (*M* = 97.6, *SD* = 36.5) and (*M* = 105, *SD* = 38.3) at the delayed posttest. Hybrid feedback group recorded a change from the pretest (*M* = 57.4, *SD* = 30.2) to the immediate posttest (*M* = 84.3, *SD* = 30.6) and to the delayed posttest (*M* = 116, *SD* = 16.2). Figure 6 presents the changes in mean scores in the three groups at the three test phases (pretest, immediate posttest and delayed posttest).

Table 6 *Descriptive Statistics for Overall Writing Scores in the Pretest, Immediate Posttest and Delayed Posttest.*

	Pretest				Immediate Posttest				Delayed Posttest			
Group	*M*	*SD*	*MD*	*IQR*	*M*	*SD*	*MD*	*IQR*	*M*	*SD*	*MD*	*IQR*
TF (*N* = 25)	47.7	34.3	44	59	85.4	35.2	87	43	86.8	43.7	90	37.2
AF (*N* = 25)	64.1	38.8	63	40	97.6	36.5	97	50	105	38.3	93.5	73
ATF (*N* = 25)	57.4	30.2	55.5	49.8	84.3	30.6	80.5	34.8	116	16.2	118	23

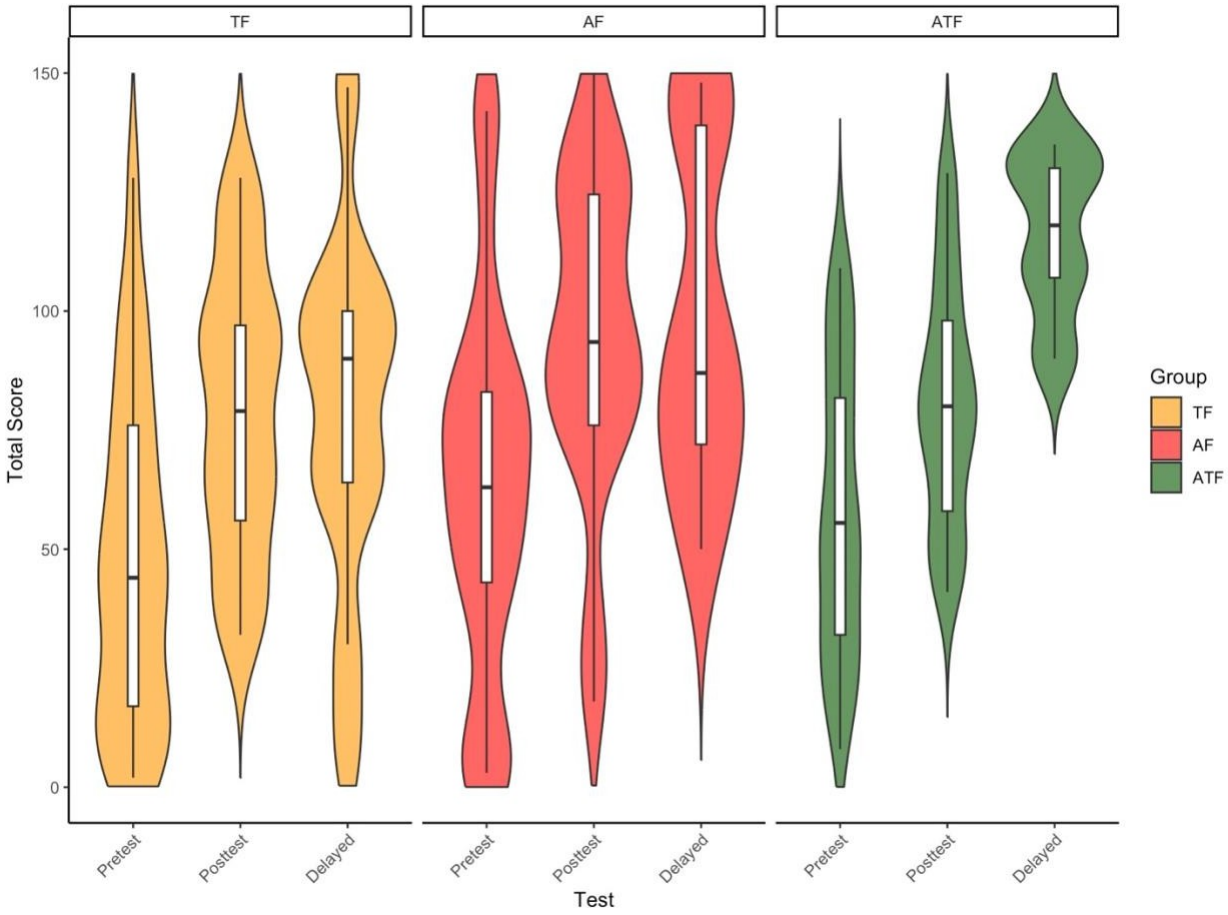


Figure 5 Group Overall Performance in L2 Writing Over Time.

The overall linear mixed-effects model revealed that the fixed effects explained a relatively small portion of the variance for the overall writing performance (marginal $R^2 = .251$), whereas the combined fixed and random effects explained a larger amount (conditional $R^2 = .778$). The comparison of the full model with all fixed effects included with the null model with fixed effects excluded confirms that the interaction between feedback type and test time (independent variables) had a significant effect on the participants' L2 writing production (dependent variable) ($\chi^2 = 8.563$, $p = .0037$).

The model estimates in Table 7 (illustrated in Figure 7) show a significant positive change within all three groups at the immediate posttest. The results for the teacher feedback group confirmed that the total writing score significantly increased from the pretest to the immediate posttest (β = 37.72 [32.95 – 42.49], t = 11.31, p < .00025). The automated feedback group recorded a significant change in score at the immediate posttest as well (β = 33.56[22.1 – 31.78], t = 13.82, p < .00025) and the hybrid feedback group model estimates confirmed a significant change at the immediate posttest (β = 26.92 [28.79 – 38.33], t = 10.86, p < .00025). Similarly, all three groups showed significant change at the delayed posttest. Teacher feedback group (β = 36.85 [30.45 – 43.24], t = 11.31, p < .00025), automated feedback group (β = 48 [41.37 – 55.21], t = 13.70, p < .00025) and the hybrid feedback group *(*β = 58.74 [51.80 – 65.67], t = 16.62, p < .00025).

Table 7 *Within-Group Comparisons of Mixed-Effects Model Outcomes for Overall Writing Production.*

	Within-group Comparisons	
Group	Immediate Posttest	Delayed Posttest
TF	β = 37.72, 95% *CI* [32.95 – 42.49], *SE* = 2.43, t(820) = 15.54, p < .00025, d = 1.1	β = 36.85, 95% *CI* [30.45 – 43.24], *SE* = 3.26, t(829) = 11.31, p < .00025, d = 1
AF	β = 33.56, 95% *CI* [22.05 – 31.78], *SE* = 2.48, t(820) = 10.86, p < .00025, d = .90	β = 48, 95% *CI* [41.37 – 55.21], *SE* = 3.53, t(832) = 13.70, p < .00025, d = 1.1
ATF	β = 26.92, 95% *CI* [28.79 – 38.33], *SE* = 2.48, t(820) = 10.86, p < .00025, d = .90	β = 58.74, 95% *CI* [51.80 – 65.67], *SE* = 3.54, t(832) = 16.62, p < .00025, d = 2.4

Notes. Alpha values adjusted to correct for four repeated tests. *p < .0125. **p < .0025. ***p < .00025.

SE = standard error; *TF* = teacher feedback group; *AF* = automated feedback group; *ATF* = automated+teacher feedback group; *d* = effect size.

Full model estimates are provided in Appendices E.1, E.2 and E.3.

The linear mixed-effects model estimates in Table 8 compare groups at the pretest, immediate posttest and delayed posttest. At the pretest, there was no significant difference between groups, as expected, and this confirms the ANOVA results obtained prior to full data analysis. Similarly, at the immediate posttest, there was no significant difference between groups. At the delayed posttest, there was a significant difference between the teacher feedback group and the automated feedback group (β = 25.75 [7.84 – 43.66], t = 2.8, p = .005), as well as between the teacher feedback group and the hybrid feedback group (β = 30.62 [12.72 – 48.52], t = 3.4, p = .001) but there was only a small difference between the automated feedback group and the hybrid feedback group (β = 4.87 [-13.19 – 22.93], t = .53, p = .597).

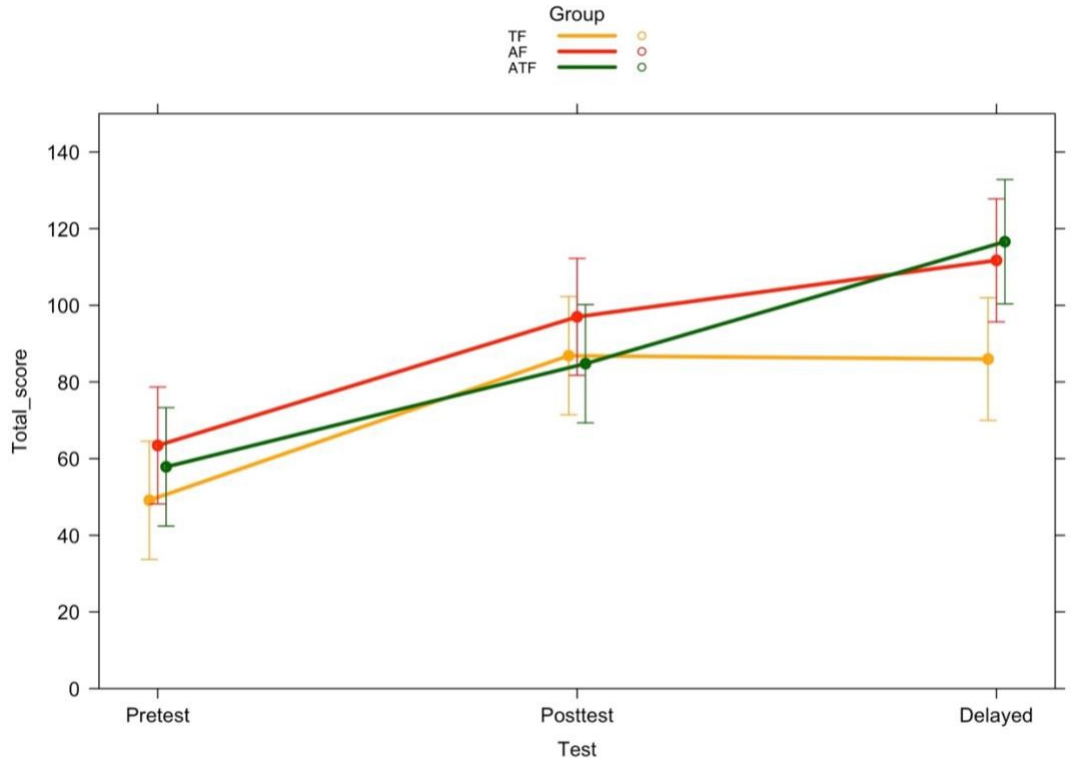


Figure 6 Plot of the Linear Mixed-effects Model Outcome of Total Writing Scores.

Table 8 *Between-Group Comparisons of Mixed-Effects Model Outcomes for Overall Writing Production.*

	Between-Group Comparisons		
Test	TF vs. AF	TF vs. ATF	AF vs ATF
Pretest	β = 14.30, 95% *CI* [-2.34 – 30.95], *SE* = 8.5, *t*(77) = 1.69, *p* = .095, *d* = .40	β = 8.72, 95% *CI* [-7.94 – 25.39], *SE* = 8.5, *t*(77) = 1.02, *p* = .307, *d* = .30	β = -5.58, 95% *CI* [-22.20 – 11.04], *SE* = 8.5, *t*(77) = -0.659, *p* = .512, *d* = .20
Immediate posttest	β = 10.14, 95% *CI* [-6.50 – 26.79], *SE* = 8.5, *t*(77) = 1.19, *p* = .235, *d* = .30	β = -2.08, 95% *CI* [-18.75 – 14.59], *SE* = 8.5, *t*(77) = -0.245, *p* = .807, *d* = .03	β = -12.22, 95% *CI* [-28.84 – 4.40], *SE* = 8.5, *t*(77) = -1.44, *p* = .153, *d* = .40
Delayed posttest	β = 25.75, 95% *CI* [7.84 – 43.66], *SE*= 9.1, *t*(103)= 2.8, *p* = .005, *d* = .40	β = 30.62, 95% *CI* [12.72 – 48.52], *SE* = 9.2, *t*(102) = 3.4, *p* = .001, *d* = .90	β = 4.87, 95% *CI* [-13.19 – 22.93], *SE* = 9.1, *t*(106) = .53, *p* = .597, *d* = .40

Notes. Alpha values adjusted to correct for four repeated tests. *p < .0125. **p < .0025. ***p < .00025.

SE = standard error; *TF* = teacher feedback group; *AF* = automated feedback group; *ATF* = automated+teacher feedback group; *d* = effect size.

Full model estimates are provided in Appendices E.1, E.2 and E.3.

Cohen's *d* effect sizes (see tables 7,8) illustrate the magnitude of the effects observed in both within-group and between-group comparisons. For the teacher feedback group, a medium effect was found for the comparison between the pretest and immediate posttest (*d* = 1.1), as well as between the pretest and delayed posttest (*d* = 1.0). Similarly, the automated feedback group showed a small effect for the pretest to immediate posttest comparison (*d* = 0.9) and a medium effect for the pretest to delayed posttest comparison (*d* = 1.1). The hybrid feedback group demonstrated small effect (*d* = .90) for the pretest to immediate posttest comparison, (*d* = 2.4) for

the pretest to delayed posttest comparison. These results suggest a slightly stronger effect of the hybrid feedback group at the delayed posttest time.

For between-group comparisons, effects were generally small or negligible at the immediate posttest ($d > .50$) for all three feedback groups. However, at the delayed posttest, a medium effect ($d = .90$) was observed when comparing the teacher feedback group to the hybrid feedback group.

In summary, the timing of the test had a significant impact on participants' overall writing scores across all three feedback conditions. This confirms that all three feedback types had a positive effect on writing performance. However, differences between feedback types were minimal, with the largest effect observed for the hybrid (automated + teacher) feedback group at the delayed posttest.

5.3 Research Question 2: Comparisons of the Effects of Feedback Condition on L2 Writing Components (Text Content, Text Organisation, Vocabulary Use, Language Use and Mechanics)

This section sets out to answer the second research question: To what extent do three different feedback conditions (teacher only, automated only, hybrid feedback) affect the different components of L2 writing examined: text content, text organisation, vocabulary use, language use and mechanics?

Initially, ANOVA was employed before conducting data analysis to ensure the comparability of the groups in each of the tested writing components prior to the experiment. Descriptive statistics are reported next to provide an overview of the data. The results for the five linear mixed-effect models and Cohen's *d* effect sizes are subsequently presented. For each model, the focus is on the fixed effect estimates which directly address the research question.

5.3.1 Text Content

The ANOVA results indicated no significant difference among the three participant groups in pretest text content scores, $F(2, 71) = 2.15$, $p = .124$.

Descriptive statistics (mean scores, standard deviation, median, and interquartile ranges) presented in Table 9 revealed an increase in mean scores across test times. For the teacher feedback group, the mean score increased from the pretest ($M = 9.9$, $SD = 7.7$) to the immediate posttest ($M = 17.2$, $SD = 7.7$) and the delayed posttest ($M = 17.7$, $SD = 8.9$). Similarly, the automated feedback group showed an increase in mean scores from the pretest ($M = 14.2$, $SD = 8.1$) to the immediate posttest ($M = 20.8$, $SD = 8.3$) and the delayed posttest ($M = 21.6$, $SD = 8.2$). For the hybrid feedback group, the mean score also increased from the pretest ($M = 13.0$,

SD = 6.8) to the immediate posttest (*M* = 18.3, *SD* = 7.0) and the delayed posttest (*M* = 25.7, *SD* = 4.0).

Table 9 *Summary of Descriptive Statistics for Text Content Scores Obtained at Pretest, Immediate Posttest and Delayed Posttest.*

	Pretest				Immediate Posttest				Delayed Posttest			
Group	*M*	*SD*	*MD*	*IQR*	*M*	*SD*	*MD*	*IQR*	*M*	*SD*	*MD*	*IQR*
TF (*N* = 25)	9.9	7.7	9.0	11.5	17.2	7.7	18.0	9.0	17.7	8.9	20.0	7.0
AF (*N* = 25)	14.2	8.1	14.0	9.0	20.8	7.6	20.0	8.0	21.6	8.2	20.0	12.3
ATF (*N* = 24)	13.0	6.8	13.0	11.3	18.3	7.0	18.0	9.3	25.7	4.0	25.0	6.8

The model estimates in Table 10 (illustrated in Figure 8) indicate a significant positive change in text content scores in all three groups at the immediate posttest. The teacher feedback group experienced a significant increase in text content scores from pretest to immediate posttest (β = 7.36 [4.45 – 10.27], *t* = 5.00, *p* < .00025). The automated feedback group also recorded a significant positive change in text content score at the immediate posttest (β = 6.56 [3.65 – 9.47], *SE* = 1.47, *t* = 4.45, *p* < .00025). Hybrid feedback group model estimates confirmed a significant increase at immediate posttest (β = 5.25 [2.28 – 8.22], *t* = 3.49, *p* < .00025). The change in text content score from the pretest to the delayed posttest was also significant for all feedback groups. The change in text content scores from the pretest to the delayed posttest was also significant for all feedback groups. The teacher feedback group showed a significant improvement (β = 6.26 [2.45 – 10.08] , *t = 3.24* , *p* < .0025), while the automated feedback

group recorded a similar increase (β = 7.64 [3.53 – 11.75], t = 3.67, p < .0025). The hybrid feedback group demonstrated the largest effect (β = 12.26 [8.13 – 16.38], t = 5.87, p < .00025).

These results indicate that all three feedback types contributed to improvements in text content scores, with the hybrid feedback group exhibiting the most pronounced gains.

Table 10 *Within-Group Comparisons of Mixed-effects Model Outcome for Text Content Scores.*

	Within-Group Comparisons	
Group	Immediate Posttest	Delayed posttest
TF	β = 7.36, 95% *CI* [4.45 – 10.27], *SE* = 1.74, t(101) = 5.00, p < .00025, d = .90	β = 6.26, 95% *CI* [2.45 – 10.08] , *SE* = 1.93, t(111) = 3.24 , p < .0025, d = .90
AF	β = 6.56, 95% *CI* [3.65 – 9.47], *SE* = 1.47, t(101) = 4.45, p < .00025, d = .84	β = 7.64, 95% *CI* [3.53 – 11.75], *SE* = 2.08, t(114) = 3.67, p < .0025, d = .90
ATF	β = 5.25, 95% *CI* [2.28 – 8.22], *SE* = 1.50, t(101) = 3.49, p < .00025, d = .76	β = 12.26, 95% *CI* [8.13 – 16.38], *SE* = 2.09, t(114) = 5.87, p < .00025, d = 2.3

Notes. Alpha values adjusted to correct for four repeated tests. *p < .0125. **p < .0025. ***p < .00025.

SE = standard error; *TF* = teacher feedback group; *AF* = automated feedback group; *ATF* = automated+teacher feedback group; d = effect size.

Full model estimates are provided in Appendices E.4, E.5 and E.6.

Referring to the content, between-group comparisons presented in Table 11 show the differences between groups at the pretest, the immediate posttest, and the delayed posttest. At the pretest, there was no significant difference between groups which was expected and confirms the ANOVA results recorded prior to the full data analysis. During the immediate posttest, there were slight but insignificant differences between groups. At the delayed posttest, there was only a significant difference between the teacher feedback group and the hybrid feedback group (β = 7.91 [2.34 – 13.48], t = 2.80, p = .005).

Table 11 *Between-Group Comparison of Mixed-effects Model Outcome for Text Content Scores.*

	Between-Group Comparisons		
Test	TF vs. AF	TF vs. ATF	AF vs ATF
Pretest	β = 2.97, 95% *CI* [-1.11 – 7.05], *SE* = 2.07, t(127) = 1.44, p = .153, d = .50	β = 1.92, 95% *CI* [-2.18 – 6.02], *SE* = 2.08, t(126) = 0.92, p = .357, d = .40	β = -1.05, 95% *CI* [-5.14 – 3.04], *SE* = 2.07, t(126) = -0.51, p = .613, d = .20
Immediate posttest	β = 2.17, 95% *CI* [-1.91 – 6.25], *SE* = 2.07, t(127) = 1.05, p = .296, d = .50	β = -0.19, 95% *CI* [-4.29 – 3.91], *SE* = 2.08, t(126) = -0.09, p = .927, d = .10	β = -2.36, 95% *CI* [-6.45 – 1.73], *SE* = 2.07, t(126) = -1.14, p = .257, d = .30
Delayed posttest	β = 4.34, 95% *CI* [-1.25 – 9.94], *SE* = 2.84, t(170) = 1.53, p = .0127, d = .50	β = 7.91, 95% *CI* [2.34 – 13.48], *SE* = 2.82, t(170) = 2.80, p = .005, d = 1.2	β = 3.57, 95% *CI* [-2.22 – 9.35], *SE* = 2.93, t(1170) = 1.22, p = .225, d = .60

Notes. Alpha values adjusted to correct for four repeated tests. *p < .0125. **p < .0025. ***p < .00025.

SE = standard error; *TF* = teacher feedback group; *AF* = automated feedback group; *ATF* = automated+teacher feedback group; d = effect size.

Full model estimates are provided in Appendices E.4, E.5 and E.6.

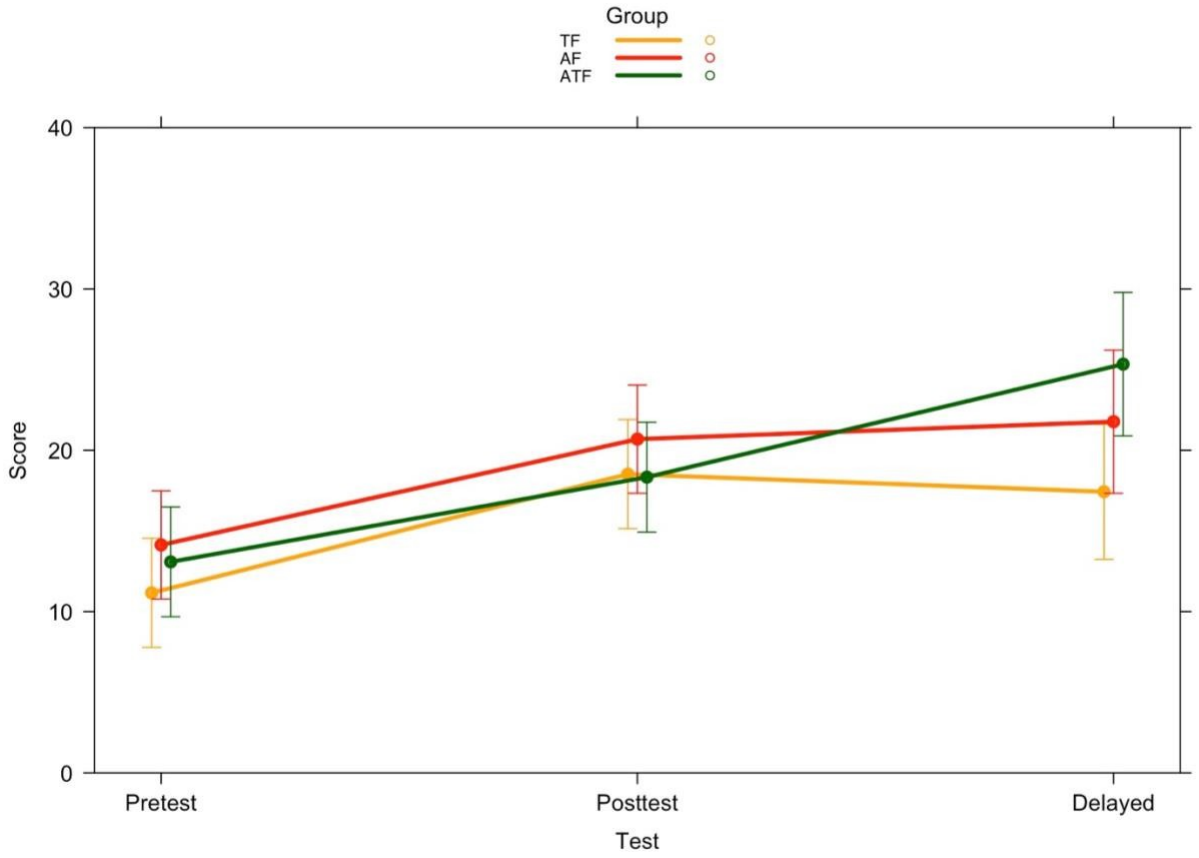


Figure 7 Plot of the Linear Mixed-effects Model Outcome for Text Content Scores.

The Cohen's *d* effect sizes (see tables 10,11) indicate that in the teacher feedback group there was a small effect between the pretest and the immediate posttest (*d* = .90). There was also a small effect between the pretest and the delayed posttest (*d* = .90). The effect of test time was also small in the automated feedback group when comparing the pretest and the immediate posttest (*d* = 0.84) and between the pretest and the delayed posttest (*d* = .90). In the hybrid feedback group, the effect sizes were slightly stronger by L2 research norms compared with other groups. Effects were small (*d* = .76) between the pretest and the immediate posttest, and large (*d* = 2.3) between the pretest and the delayed posttest.

Effect sizes for between-group comparisons of text content scores indicate that there was only a small effect between the teacher feedback group and the automated feedback group (d = .50), while the effect between the other groups at the immediate posttest was negligible (d < .50). At the delayed posttest, the effect was small (d = .50) between the teacher feedback group and the automated feedback group, large (d = 1.2) between the teacher feedback group and the hybrid feedback group and small (d = .60) between the automated feedback group and the hybrid feedback group.

Overall, there was an increase in text content scores for L2 participants in the three feedback groups. The strongest effect was observed in the comparison between the teacher feedback group and the hybrid feedback group during the delayed posttest.

5.3.2 Text Organisation

The results of the ANOVA test revealed that there was no significant difference among the three groups of pretest text organisation scores, F (2,71) = 1.368, p = .261.

Descriptive statistics for text organization scores in Table 12 demonstrate the change in scores across test times. The teacher feedback group had a mean score of (M = 8.9, SD = 8.1) at the pretest, which increased to (M = 15.7, SD = 8.1) at the immediate posttest and (M = 17.5, SD = 8.9) at the delayed posttest. Similarly, the automated feedback group had a mean score of (M = 12.2, SD = 8.6) at the pretest, increasing to (M = 20.8, SD = 8.3) at the immediate posttest and (M = 21.1, SD = 8.6) at the delayed posttest. The hybrid feedback group showed an increase from (M = 12.1, SD = 7.2) at the pretest to (M = 17.6, SD = 7.0) at the immediate posttest and (M = 24.4, SD = 4.0) at the delayed posttest.

Table 12 *Summary of Descriptive Statistics for Text Organisation Scores at Pretest, Immediate Posttest and Delayed Posttest.*

	Pretest				Immediate Posttest				Delayed Posttest			
Group	*M*	*SD*	*MD*	*IQR*	*M*	*SD*	*MD*	*IQR*	*M*	*SD*	*MD*	*IQR*
TF (N = 25)	8.9	8.1	7.5	13.3	15.7	8.1	15.5	9.3	17.5	9.3	20.0	9.3
AF (N = 25)	12.2	8.6	10.0	10.0	20.8	8.3	20.0	12.0	21.1	8.6	20.0	14.8
ATF (N = 24)	12.1	7.2	12.0	10.8	17.6	7.0	17.5	10.3	24.4	4.0	24.5	4.8

The model estimates in Table 13 (illustrated in Figure 9) indicate a significant positive change in text organisation scores in all three groups at the immediate posttest phase. The teacher feedback group recorded significantly increased text organisation scores from pretest to immediate posttest (β = 6.88 [3.70 – 10.06], t = 4.28, p < .00025), and from the pretest to the delayed posttest

Similarly, the automated feedback group demonstrated a significant improvement in text organization scores at the immediate posttest (β = 8.52 [5.34 – 11.7], t = 5.29, p < .00025), and at the delayed posttest (β = 9.83 [5.36 – 14.30], t = 4.34, p < .0025). The hybrid feedback group also exhibited significant increases at the immediate posttest (β = 5.50 [2.26 – 8.74], t = 3.35, p < .0025) and at the delayed posttest (β = 12.24 [7.76 – 16.73], t = 5.39, p < .00025)

For between-group comparisons, there was no significant difference in text organization scores between the groups across test times (see Table 14).

Table 13 *Within-Group Comparisons of Mixed-effects Model Outcome for Text Organisation Scores.*

	Within-Group Comparisons	
Group	Immediate Posttest	Delayed Posttest
TF	β = 6.88, 95% *CI* [3.70 – 10.06], *SE* = 1.61, *t*(101) = 4.28, $p < .00025$, *d* = .80	β = 7.45, 95% *CI* [3.28 – 11.60], *SE* = 2.11, *t*(113) = 3.53, $p < .0025$, *d* = 1.0
AF	β = 8.52, 95% *CI* [5.34 – 11.7], *SE* = 1.61, *t*(101) = 5.29, $p < .00025$, *d* = 1.0	β = 9.83, 95% *CI* [5.36 – 14.30], *SE* = 2.26, *t*(116) = 4.34, $p < .0025$, *d* = 1.0
ATF	β = 5.50, 95% *CI* [2.26 – 8.74], *SE* = 1.64, *t*(101) = 3.35, $p < .0025$, *d* = .77	β = 12.24, 95% *CI* [7.76 – 16.73], *SE* = 2.27, *t*(116) = 5.39, $p < .00025$, *d* = 2.1

Notes. Alpha values adjusted to correct for four repeated tests. *$p < .0125$. **$p < .0025$. ***$p < .00025$.

SE = standard error; *TF* = teacher feedback group; *AF* = automated feedback group; *ATF* = automated+teacher feedback group; *d* = effect size.

Full model estimates are provided in Appendices E.7, E.8 and E.9.

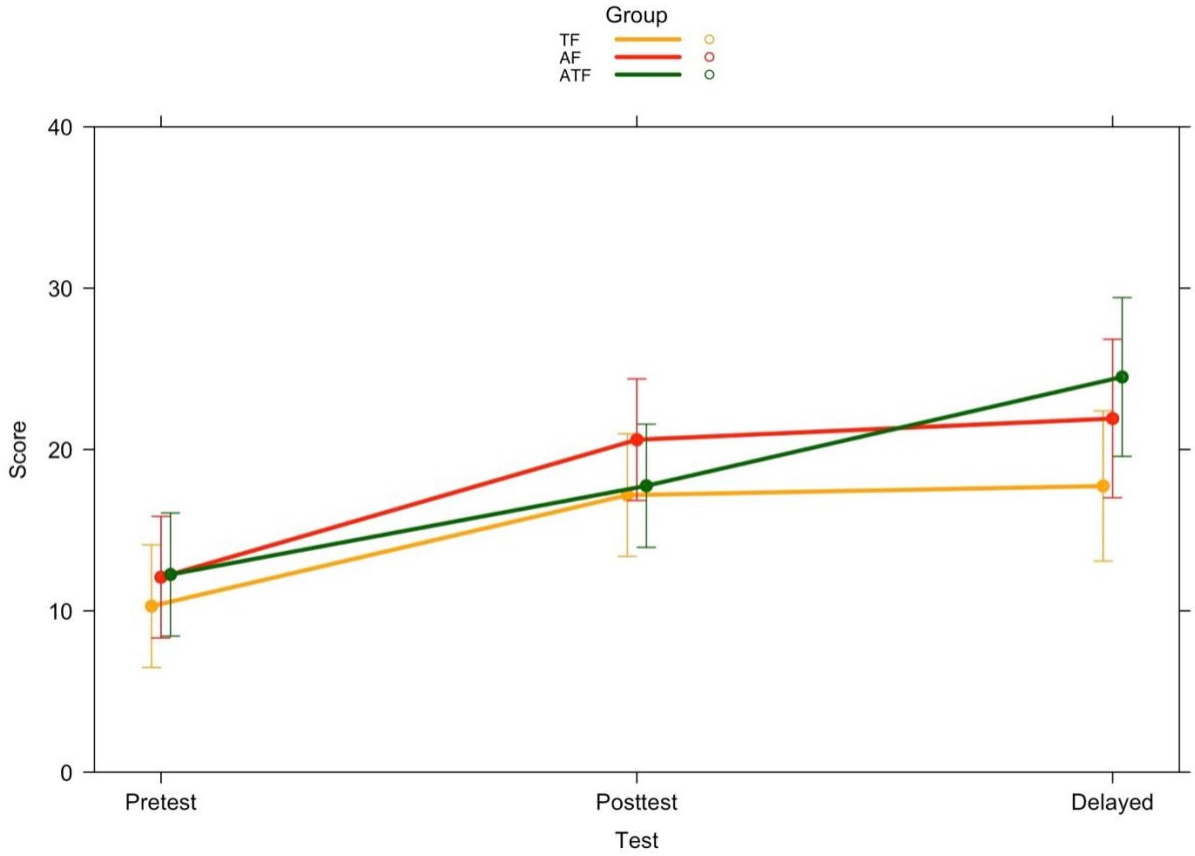


Figure 8 Plot of the Linear Mixed-effects Model Outcome for Text Organisation Scores.

Table 14 *Between-Group Comparison of Mixed-effects Model Outcome for Text Organisation Scores.*

	Between-Group Comparisons		
Test	TF vs. AF	TF vs. ATF	AF vs ATF
Pretest	β = 1.79, 95% *CI* [-2.48 – 6.07], *SE* = 2.17, *t*(132) = 0.83, *p* = .472, *d* = .40	β = 1.96, 95% *CI* [-2.33 – 6.26], *SE* = 2.18, *t*(132) = 0.90, *p* = .369, *d* = .40	β = 0.17, 95% *CI* [-4.12 – 4.45], *SE*= 2.17, *t*(132)= -0.83, *p*=.938, *d* = .01
Immediate posttest	β = 3.43, 95% *CI* [-0.84 – 7.71], *SE* = 2.17, *t*(132) = 1.59, *p* = .115, *d* = .60	β = 0.58, 95% *CI* [-3.71 – 4.88], *SE* = 2.18, *t*(132) = 0.27, *p* = .789, *d* = .30	β = -2.85, 95% *CI* [-7.14 – 1.43], *SE* = 2.17, *t*(132) = -1.31, *p* = .191, *d* = .40
Delayed posttest	β = 4.18, 95% *CI* [-1.78 – 10.15], *SE* = 3.03, *t*(171) = 1.38, *p* = .167, *d* = .40	β = 6.76, 95% *CI* [0.83 – 12.69], *SE*= 3.0, *t*(169)= 2.3, *p*=.026, *d* = 1.0	β = 2.58, 95% *CI* [-3.58 – 8.74], *SE* = 3.12, *t*(170) = 0.83, *p* = .41, *d* = .50

Notes. Alpha values adjusted to correct for four repeated tests. *p < .0125. **p < .0025. ***p < .00025.

SE = standard error; *TF* = teacher feedback group; *AF* = automated feedback group; *ATF* = automated+teacher feedback group; *d* = effect size.

Full model estimates are provided in Appendices E.7, E.8 and E.9.

Cohen's *d* effect sizes in (see tables 13,14) show that in the teacher feedback group there was a small effect between the pretest and the immediate posttest (*d* = .80), and a medium effect between the pretest and the delayed posttest (*d* = 1.0) regarding text organisation scores. The effect of test time was medium within the automated feedback group when comparing the pretest and the immediate posttest, and between the pretest and the delayed posttest (*d* = 1.0). In

the hybrid feedback group, effect sizes were slightly stronger by L2 research norms compared with other groups. Effects were small (d = .77) between the pretest and the immediate posttest, and larger (d = 2.1) between the pretest and the delayed posttest.

Effect sizes for between-group comparisons of text organisation scores indicate that at the immediate posttest there was also a small effect between the teacher feedback group and the automated feedback group (d = .60) and between the automated feedback group and the hybrid feedback group (d = .40) and a negligible effect between the teacher feedback group and the hybrid feedback group to terms of text organisation scores. At the delayed posttest, the effect was small (d = .40) for the comparison between the teacher feedback group and the automated feedback group, large (d = 1.1) between the teacher feedback group and the hybrid feedback group, and small (d = .50) between the automated feedback group and the hybrid feedback group.

5.3.3 Vocabulary Use

The results of the ANOVA test revealed that there was no significant difference among the three groups of participants in terms of the pretest vocabulary scores, F (2,71) = 2.485, p = .090. Descriptive statistics for the writing outcome regarding vocabulary use (see Table 15) indicated that in the teacher feedback group there was an increase from the pretests (M = 8.8, SD = 7.3) to the immediate posttest (M = 16.8, SD = 7.9) and the delayed posttest (M = 17.9, SD = 8.6). The automated feedback group indicated an increase in vocabulary use scores from the pretest (M = 13.1, SD = 7.6) to the immediate posttest (M = 20.5, SD = 7.9) and the delayed posttest (M = 21.3, SD = 7.6). The hybrid feedback group also improved their performance regarding vocabulary use from (M = 12.1, SD = 6.4) to the immediate posttest (M = 17.2 , SD = 5.7) and the delayed posttest (M = 25.7, SD = 3.8).

Table 15 *Summary of Descriptive Statistics for Vocabulary Use Scores at Pretest, Immediate Posttest and Delayed Posttest.*

	Pretest				Immediate Posttest				Delayed Posttest			
Group	*M*	*SD*	*MD*	*IQR*	*M*	*SD*	*MD*	*IQR*	*M*	*SD*	*MD*	*IQR*
TF (N = 25)	8.8	7.3	8.0	9.3	16.8	7.9	18.5	13.0	17.9	8.6	20.0	7.3
AF (N = 25)	13.1	7.6	15.0	9.0	20.5	7.9	20.0	10.0	21.3	7.6	20.0	11.5
ATF (N = 24)	12.1	6.4	11.5	10.3	17.2	5.7	17.0	7.3	25.7	3.8	26.0	5.3

The summary of model estimates in Table 16 (illustrated in Figure 10) indicates a significant positive change in vocabulary use in all three groups from the pretest to the immediate posttest phase. The teacher feedback group recorded a significant increase at the immediate posttest ($\beta = 8.04$ [5.38 – 10.70], $t = 5.97$, $p < .00025$) and the automated feedback group also demonstrated a significant change at immediate posttest ($\beta = 7.40$ [4.74 – 10.06, $t = 5.49$, $p < .00025$). Hybrid feedback group model estimates confirmed a significant increase at immediate posttest ($\beta = 5.08$ [2.37 – 7.80], $t = 3.70$, $p < .00025$).

At the delayed posttest, vocabulary use scores significantly changed across all three feedback groups. The teacher feedback group showed an increase from the pretest to the delayed posttest ($\beta = 8.14$ [5.38 – 10.70], $t = 4.59$, $p < .00025$). The automated feedback group also recorded a change in scores ($\beta = 9.2$ [5.43 – 12.96], $t = 4.82$, $p < .00025$). The hybrid feedback group had the largest change between the pretest and delayed posttest ($\beta = 13.67$ [9.89 – 17.45], $t = 7.14$, $p < .00025$).

Table 16 *Within-Group Comparisons of Mixed-effects Model Outcome for Vocabulary Use Scores.*

	Within-Group Comparisons	
Group	Immediate Posttest	Delayed Posttest
TF	β = 8.04, 95% *CI* [5.38 – 10.70], *SE* = 1.35, $t(102) = 5.97$, $p <$ .00025, $d = 1.0$	β = 8.14, 95% *CI* [5.38 – 10.70], *SE* = 1.77, $t(111) = 4.59$, $p <$ .00025, $d = 1.1$
AF	β = 7.40, 95% *CI* [4.74 – 10.06], *SE* = 1.35, $t(102) = 5.49$, $p <$ .00025, $d = 1.0$	β = 9.20, 95% *CI* [5.43 – 12.96], *SE* = 1.91, $t(114) = 4.82$, $p <$ .00025, $d = 1.1$
ATF	β = 5.08, 95% *CI* [2.37 – 7.80], *SE* = 1.37, $t(102) = 3.70$, $p <$ .00025, $d = .80$	β = 13.67, 95% *CI* [9.89 – 17.45], *SE* = 1.91, $t(113) = 7.14$, $p <$ .00025, $d = 2.6$

Notes. Alpha values adjusted to correct for four repeated tests. *$p < .0125$. **$p < .0025$. ***$p < .00025$.

SE = standard error; *TF* = teacher feedback group; *AF* = automated feedback group; *ATF* = automated+teacher feedback group; *d* = effect size.

Full model estimates are provided in Appendices E.10, E.11 and E.12.

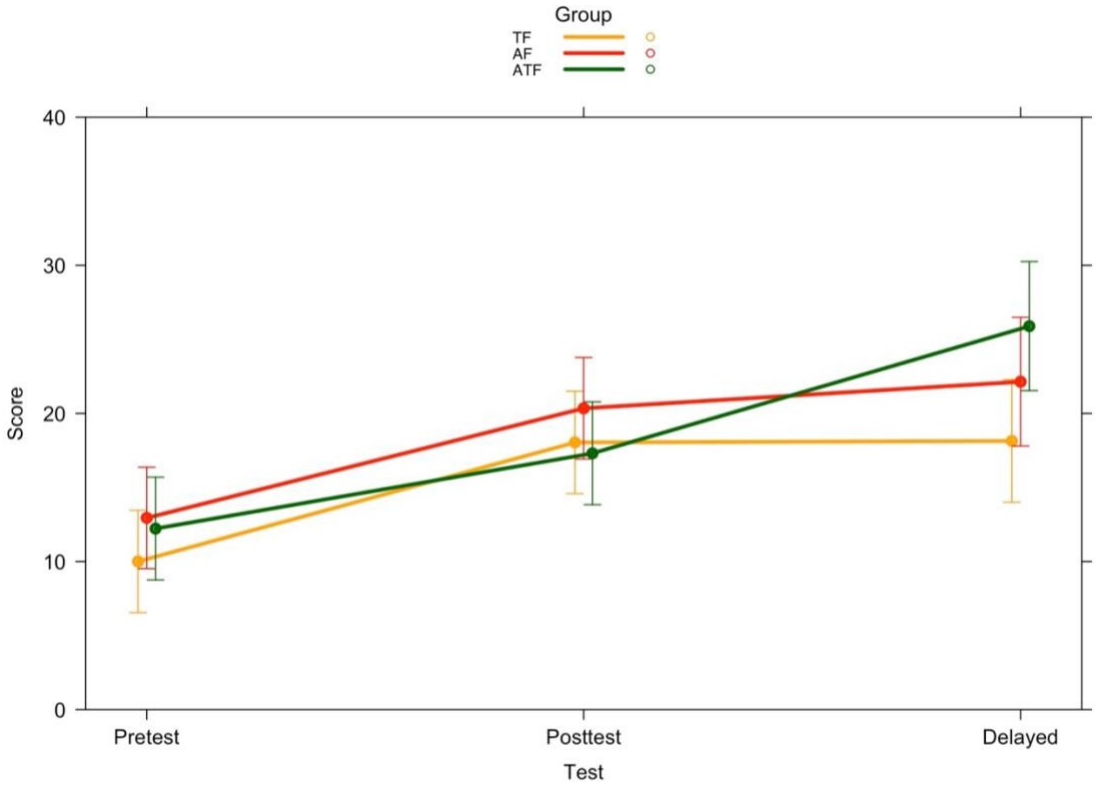


Figure 9 Plot of the Linear Mixed-effects Model Outcome of Vocabulary Use Scores.

Between-group comparisons model estimates (see Table 17) revealed that there were no significant differences at pretest which again confirmed the previous ANOVA results. Also, at the immediate posttest, no significant differences between groups were observed. However, at the delayed posttest there was only a significant difference between the teacher feedback group and the hybrid feedback group (β = 7.75 [2.54 – 12.96], t = 2.94, p = .0038*).

Table 17 *Between-Group Comparisons of Mixed-effects Model Outcome for Vocabulary Use Scores.*

	Between-Group Comparisons		
Test	TF vs. AF	TF vs. ATF	AF vs ATF
Pretest	β = 2.94, 95% *CI* [-0.93 – 6.81], *SE* = 1.96, t(123) = 1.50, p = .136, d = .60	β = 2.22, 95% *CI* [-1.66 – 6.11], *SE* = 1.97, t(122) = 1.13 , p = .261, d = .50	β = -0.72, 95% *CI* [-4.60 – 3.16], *SE* = 1.96, t(122) = -0.37, p = .714, d = .10
Immediate posttest	β = 2.30, 95% *CI* [-1.57 – 6.17], *SE* = 1.96, t(123) = 1.17, p = .242, d = .50	β = -0.73, 95% *CI* [-4.62 – 3.1], *SE* = 1.97, t(122) = -0.37, p = .709, d = .10	β = -3.04, 95% *CI* [-6.91 – 0.84], *SE* = 1.96, t(122) = -1.55, p = .124, d = .50
Delayed posttest	β = 4.00, 95% *CI* [-1.23 – 9.23], *SE* = 2.65, t(170) = 1.51, p = .133, d = .40	β = 7.75, 95% *CI* [2.54 – 12.96], *SE* = 2.64, t(168) = 2.94, p = .0038, d = 1.2	β = 3.75, 95% *CI* [-1.65 – 9.15], *SE* = 2.73, t(170) = 1.37, p = .171, d = .70

Notes. Alpha values adjusted to correct for four repeated tests. *p < .0125. **p < .0025. ***p < .00025.

SE = standard error; *TF* = teacher feedback group; *AF* = automated feedback group; *ATF* = automated+teacher feedback group; d = effect size.

Full model estimates are provided in Appendices E.10, E.11 and E.12.

Cohen's *d* effect sizes regarding vocabulary scores in (see tables 16,17) show that within the teacher feedback group there was a medium effect between the pretest and the immediate posttest ($d = 1.1$), and a medium effect between the pretest and the delayed posttest ($d = 1.1$).

Within the automated feedback group, the effect of test time was medium in the comparison between the pretest and the immediate posttest ($d = 1.0$) and also between the pretest and the delayed posttest ($d = 1.1$)

In the hybrid feedback group, the effect size was small ($d = .80$) between the pretest and the immediate posttest, and larger ($d = 2.6$) between the pretest and the delayed posttest.

Effect sizes for between-group comparisons of vocabulary use indicate that at the immediate posttest there was a small effect between the teacher feedback group and automated feedback group ($d = .50$), and a small effect between the automated feedback group and the hybrid feedback group ($d = .50$). At the delayed posttest, the effect was small ($d = .40$) between the teacher feedback group and the automated feedback group, large ($d = 1.2$) between the teacher feedback group and the hybrid feedback group and medium ($d = .70$) between the automated feedback group and the hybrid feedback group.

5.3.4 Language Use

The results of the ANOVA test revealed that there was no significant difference between the three groups of participants in terms of the pretest language use scores, F (2,71) = 2.318, p = .105.

Descriptive statistics for the change in language use over time also indicated a consistent increase in all groups. The teacher feedback group scored (M = 8.8, SD = 8) at the pretest, (M = 17.4, SD = 9.1) at the immediate posttest and (M = 19.2, SD = 9.7) at the delayed posttest. The automated feedback group changed scores from the pretest (M = 13.7, SD = 8.7) to the immediate posttest (M = 21.4, SD = 8.8) and the delayed posttest (M = 22.0, SD = 7.7). The hybrid automated+teacher feedback group also changed language use scores from the pretest (M = 11.7, SD = 7.5) to the immediate posttest (M = 17.6, SD = 7.0) and the delayed posttest (M = 25.9, SD = 4.2) (see Table 18).

Table 18 *Summary of Descriptive Statistics for Language Use Scores Recorded in Pretest, Immediate Posttest and Delayed Posttest.*

	Pretest				Immediate Posttest				Delayed Posttest			
Group	*M*	*SD*	*MD*	*IQR*	*M*	*SD*	*MD*	*IQR*	*M*	*SD*	*MD*	*IQR*
TF (N = 25)	8.8	8.0	6.5	10.3	17.4	9.1	16.5	15.0	19.2	9.7	20.0	11.3
AF (N = 25)	13.7	8.7	14.0	12.0	21.4	8.8	22.0	13.0	22.0	7.7	20.0	13.3
ATF (N = 24)	11.7	7.5	10.0	12.0	17.6	7.0	17.0	11.0	25.9	4.2	26.0	6.0

The summary of model estimates in Table 19 (illustrated in Figure 11) indicates a significant change in language use scores across all three feedback groups at both the immediate posttest and delayed posttest. At the immediate posttest, the teacher feedback group had (β = 8.80 [5.77 – 11.8], t = 5.74, p < .00025), the automated feedback group had (β = 7.68 [4.65 – 10.71], t = 5.01, p < .00025), and the hybrid feedback group had (β = 5.88 [2.79 – 8.96], t = 3.76, p < .00025). At the delayed posttest, the teacher feedback group had (β = 8.93 [4.95 – 12.92], t = 4.43, p < .00025), the automated feedback group had (β = 9.73 [5.44 – 14.02], t = 4.48, p < .00025), and the hybrid feedback group had (β = 14.52 [10.21 – 18.82], t = 6.66, p < .00025).

For between-group comparisons, the model estimates (see Table 20) indicated no significant differences between groups at the pretest, aligning with the ANOVA results. Additionally, no significant differences were observed between groups at the immediate posttest or delayed posttest in relation to language use scores.

Table 19 *Within-Group Comparisons of Mixed-effects Model Outcome for Language Use Scores.*

	Within-Group Comparisons	
Group	Immediate Posttest	Delayed posttest
TF	β = 8.80, 95% *CI* [5.77 – 11.8] , *SE* = 1.53, *t*(101) = 5.74, *p* < .00025, *d* = 1.0	β = 8.93, 95% *CI* [4.95 – 12.92], *SE* = 2.02, *t*(111) = 4.43, *p* < .00025, *d* = 1.2
AF	β = 7.68, 95% *CI* [4.65 – 10.71], *SE* = 1.53, *t*(111) = 5.01, *p* < .00025, *d* = .90	β = 9.73, 95% *CI* [5.44 – 14.02], *SE* = 2.17, *t*(113) = 4.48, *p* < .00025, *d* = 1.0
ATF	β = 5.88, 95% *CI* [2.79 – 8.96], *SE* = 1.57, *t*(101) = 3.76, *p* < .00025, *d* = .80	β = 14.52, 95% *CI* [10.21 – 18.82], *SE* = 2.18, *t*(113) = 6.66, *p* < .00025, *d* = 2.3

Notes. Alpha values adjusted to correct for four repeated tests. **p* < .0125. ***p* < .0025. ****p* < .00025.

SE = standard error; *TF* = teacher feedback group; *AF* = automated feedback group; *ATF* = automated+teacher feedback group; *d* = effect size.

Full model estimates are provided in Appendices E.13, E.14 and E.15.

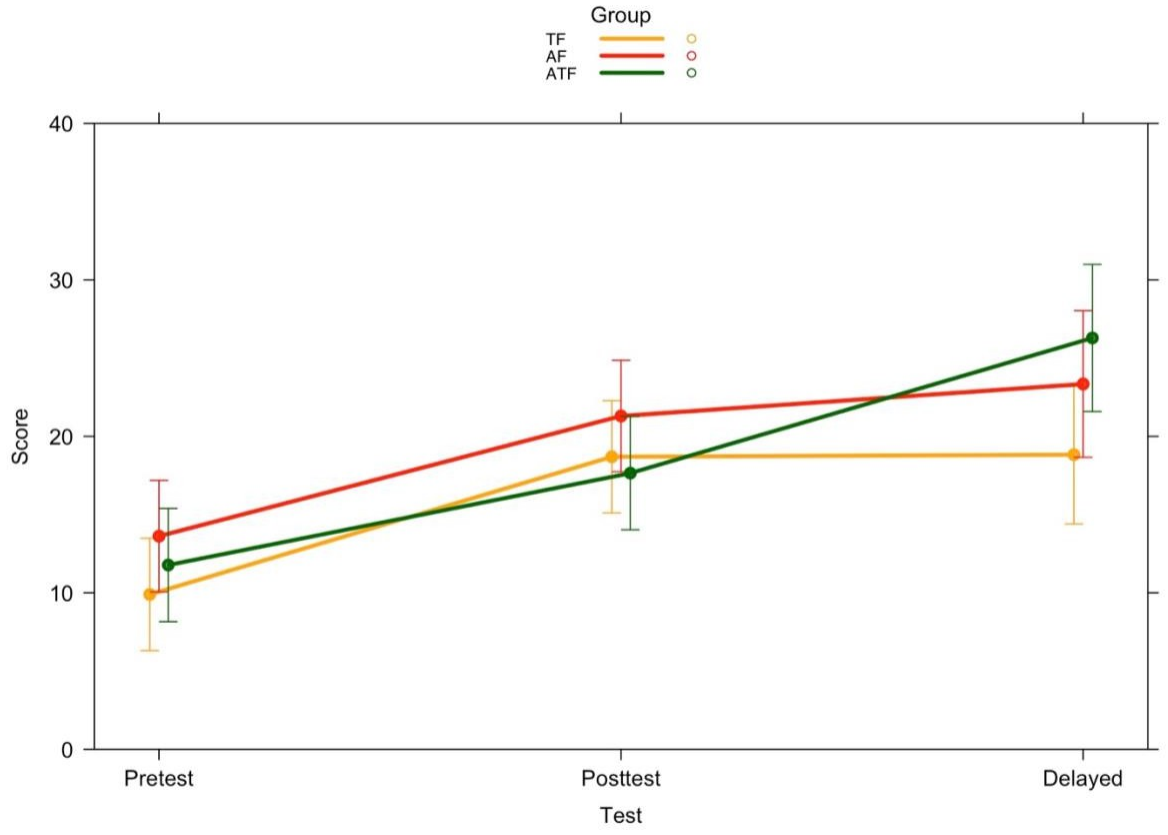


Figure 10 Plot of the Linear Mixed-effects Model Outcome of Language Use Scores.

Table 20 *Between-Group Comparisons of Mixed-effects Model Outcomes for Language Use Scores.*

	Between-Group Comparisons		
Test	TF vs. AF	TF vs. ATF	AF vs ATF
Pretest	β = 3.73, 95% *CI* [-0.72 – 8.17], *SE* = 2.25, *t*(122) = 1.66, *p* = .100, *d* = .60	β = 1.88, 95% *CI* [-2.59 – 6.34], *SE* = 2.26, *t*(121) = 0.83, *p* = .407, *d* = .40	β = -1.85, 95% *CI* [-6.30 – 2.61], *SE* = 2.26, *t*(121) = -0.82, *p*=.415, *d* = .20
Immediate posttest	β = 2.61, 95% *CI* [-1.84 – 7.05], *SE* = 2.25, *t*(122) = 1.16, *p* = .249, *d* = .40	β = -1.05, 95% *CI* [-5.51 – 3.42], *SE* = 2.26, *t*(121) = -0.46, *p* = .644, *d* = .02	β = -3.65, 95% *CI* [-8.11 – 0.81], *SE* = 2.26, *t*(121) = -1.62, *p* = .108, *d* = .50
Delayed posttest	β = 4.52, 95% *CI* [-1.47 – 10.50], *SE* = 3.03, *t*(170) = 1.49, *p* = .138, *d* = .30	β = 7.46, 95% *CI* [1.50 – 13.43], *SE* = 3.02, *t*(168) = 2.47, *p* = .015, *d* = .90	β = 2.94, 95% *CI* [-3.24 – 9.12], *SE* = 3.13, *t*(170) = 0.94, *p* = .348, *d* = .60

Notes. Alpha values adjusted to correct for four repeated tests. *p < .0125. **p < .0025. ***p < .00025.

SE = standard error; *TF* = teacher feedback group; *AF* = automated feedback group; *ATF* = automated+teacher feedback group; *d* = effect size.

Full model estimates are provided in Appendices E.13, E.14 and E.15.

Cohen's *d* effect sizes for language use scores (see tables 19, 20) indicate variation in effect sizes across test times within and between feedback groups.

For within-group comparisons, the teacher feedback group showed a medium effect between the pretest and immediate posttest (*d* = 1.0) and a medium effect between the pretest and delayed posttest (*d* = 1.2). Within the automated feedback group, the effect of test time was small but close to medium effect for the pretest to immediate posttest comparison (*d* = .90) and medium for the pretest to delayed posttest comparison (*d* = 1.0). The hybrid feedback group

exhibited small effects between the pretest and immediate posttest (d = .80), and larger effect between the pretest and delayed posttest (d = 2.3).

Effect sizes for between-group comparisons of language use scores indicate small to negligible differences at the immediate posttest. A small effect was observed between the teacher feedback group and the automated feedback group (d = .40), while the difference between the teacher feedback group and the hybrid feedback group was negligible. A small effect was also found between the automated feedback group and the hybrid feedback group (d = .50).

At the delayed posttest, the effect size remained negligible for the comparison between the teacher feedback group and the automated feedback group. A medium effect was observed in the comparison between the teacher feedback group and the hybrid feedback group (d = .90), while a small effect was found in the comparison between the automated feedback group and the hybrid feedback group (d = .60).

5.3.5 Mechanics

The ANOVA results revealed that there was no significant difference among the three groups of participants at the pretest in terms of mechanics scores, $F(2, 71) = 2.379$, $p = .099$. Mechanics descriptive statistics for the change in scores over time (see Table 21) indicate that the teacher feedback group scored ($M = 6.9$, $SD = 6.2$) at the pretest, ($M = 13.2$, $SD = 7.8$) at the immediate posttest and (M = 14.5, SD = 10.9) at the delayed posttest. The automated feedback group scored ($M = 8.0$, $SD = 7.9$) at the pretest, ($M = 14.2$, $SD = 7.9$) at the immediate posttest, and ($M = 18.9$, $SD = 9.4$) at the delayed posttest. The hybrid automated+teacher feedback group scored ($M = 8.5$, $SD = 4.4$) at the pretest, ($M = 13.7$, $SD = 7.1$) at the immediate posttest and ($M = 14.4$, $SD = 6.2$) at the delayed posttest.

Table 21 *Summary of Descriptive Statistics for Mechanics Scores Recorded in Pretest, Immediate Posttest and Delayed Posttest.*

	Pretest				Immediate Posttest				Delayed Posttest			
Group	*M*	*SD*	*MD*	*IQR*	*M*	*SD*	*MD*	*IQR*	*M*	*SD*	*MD*	*IQR*
TF ($N = 25$)	6.9	6.2	5.5	9.3	13.2	7.8	10.0	8.5	14.5	10.9	10.0	4.0
AF ($N = 25$)	10.8	7.9	10.0	7.0	14.2	7.9	10.0	10.0	18.9	9.4	16.0	17.0
ATF ($N = 24$)	8.5	4.4	8.5	5.3	13.7	7.1	10.0	5.8	14.4	6.2	11.0	7.5

The summary of model estimates in Table 22 (illustrated in Figure 12) indicates a significant positive change in mechanics scores within the teacher feedback group (β = 6.73 [3.45 – 10.0], t = 4.05, p < .00025) and the hybrid feedback group (β = 5.47 [2.06 – 8.87], t = 3.17, p < .0025) at immediate posttest but only a small increase for the automated feedback group (β = 3.55 [0.26 – 6.84], t = 2.13, p = .035). At the delayed posttest, the teacher feedback group had (β = 7.48 [3.22 – 11.73], t = 3.47, p < .0025), the automated feedback group had (β = 8.75 [4.18 – 13.32], t = 3.78, p < .00025), and the hybrid feedback group had (β = 6.81 [2.19 – 11.42], t = 2.91, p < .0125).

In terms of between-group comparisons, model estimates (see Table 23) revealed that there were no significant differences between groups at pretest, which again confirmed the ANOVA results obtained earlier. There were no significant differences at the immediate posttest and the delayed posttest times between groups regarding the mechanics scores.

Table 22 *Within-Group Comparisons of Mixed-effects Model Outcomes for Mechanics Scores.*

	Within-Group Comparisons	
Group	Immediate Posttest	Delayed Posttest
TF	β = 6.73, 95% *CI* [3.45 – 10.0], *SE* = 1.66, $t(101) = 4.05$, $p < .00025$, $d = .90$	β = 7.48, 95% *CI* [3.22 – 11.73], *SE* = 2.16, $t(115) = 3.47$, $p < .0025$, $d = .90$
AF	β = 3.55, 95% *CI* [0.26 – 6.84], *SE* = 1.67, $t(101) = 2.13$, $p = .035$, $d = .40$	β = 8.75, 95% *CI* [4.18 – 13.32], *SE* = 2.32, $t(120) = 3.78$, $p < .00025$, $d = .90$
ATF	β = 5.47, 95% *CI* [2.06 – 8.87], *SE* = 1.73, $t(86) = 3.17$, $p < .0025$, $d = .90$	β = 6.81, 95% *CI* [2.19 – 11.42], *SE* = 2.34, $t(116) = 2.91$, $p < .0125$, $d = 1.1$

Notes. Alpha values adjusted to correct for four repeated tests. *$p < .0125$. **$p < .0025$. ***$p < .00025$.

SE = standard error; *TF* = teacher feedback group; *AF* = automated feedback group; *ATF* = automated+teacher feedback group; *d* = effect size.

Full model estimates are provided in Appendices E.16, E.17 and E.18.

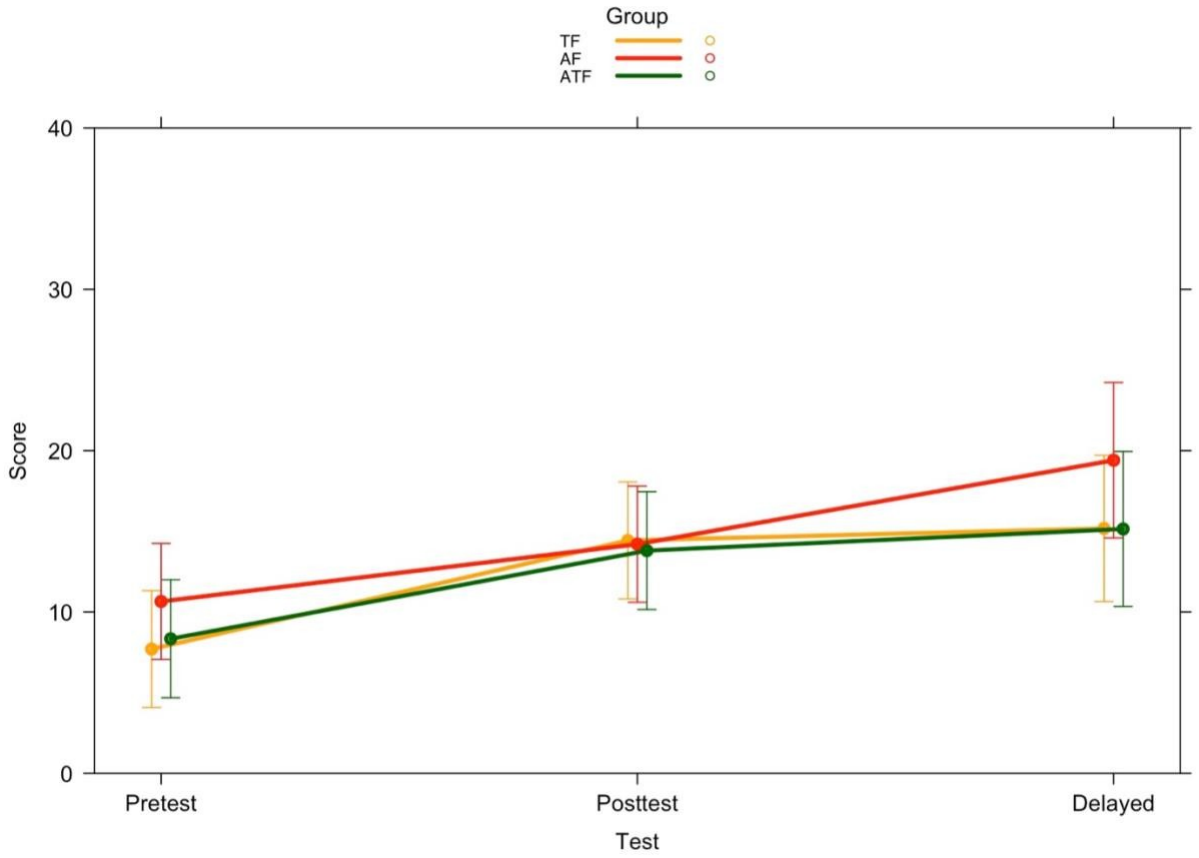


Figure 11 Plot of the Linear Mixed-effects Model Outcome of Mechanics Scores.

Table 23 *Between-Group Comparisons of Mixed-effects Model Outcomes for Mechanics Scores.*

	Between-Group Comparisons		
Test	TF vs. AF	TF vs. ATF	AF vs ATF
Pretest	β = 2.95, 95% *CI* [-1.09 – 7.00], *SE* = 2.05, *t*(144) = 1.44, *p* = .152, *d* = .50	β = .64, 95% *CI* [-3.43 – 4.71], *SE* = 2.06, *t*(145) = .31 , *p* = .758, *d* = .30	β = -2.32, 95% *CI* [-6.38 – 1.75], *SE* = 2.06, *t*(145) = 1.12, *p* = .263, *d* = .40
Immediate posttest	β = -0.23, 95% *CI* [-4.28 – 3.82], *SE* = 2.05, *t*(145) = -0.11, *p* = .910, *d* = .10	β = -0.63, 95% *CI* [-4.69 – 3.44], *SE* = 2.06, *t*(145) = -0.31, *p* = .760, *d* = .10	β = -0.40, 95% *CI* [-4.47 – 3.68], *SE* = 2.06, *t*(145) = .19, *p* = .874, *d* = .10
Delayed posttest	β = 4.23, 95% *CI* [-1.60 – 10.05], *SE* = 2.95, *t*(169) = 1.43, *p* = .154, *d* = .40	β = -0.03, 95% *CI* [-5.83 – 5.77], *SE* = 2.94, *t*(169) = .01, *p* = .991, *d* = .01	β = -4.26, 95% *CI* [-10.33 – 1.81], *SE* = 3.07, *t*(167) = -1.39, *p* = .168, *d* = .60

Notes. Alpha values adjusted to correct for four repeated tests. *p < .0125. **p < .0025. ***p < .00025.

SE = standard error; *TF* = teacher feedback group; *AF* = automated feedback group; *ATF* = automated+teacher feedback group; *d* = effect size.

Full model estimates are provided in Appendices E.16, E.17 and E.18.

Cohen's *d* effect sizes for mechanics scores (see tables 22, 23), which indicate that within-group comparisons across test times generally yielded small effect sizes (d < .90). A medium effect was only observed in the hybrid feedback group in comparisons between the pretest and the delayed posttest mechanics scores (d = 1.1).

Effect sizes for between-group comparisons at the three test times were mainly negligible. However, there was a small effect at the delayed posttest for the comparison between the teacher feedback group and the automated feedback group (d = .40) and also between the automated feedback group and the hybrid feedback group (d = .60).

All of the participants in the three feedback groups demonstrated a positive change in scores across all components (see Figures 13, 14, 15, 16) over time. At the immediate posttest, the automated feedback group outperformed the other two groups. However, at the delayed posttest, the hybrid feedback group surpassed the other groups in all writing components except for mechanics, where the automated feedback group recorded the largest increase in scores. The hybrid feedback group consistently demonstrated a marked reduction in standard deviation compared to the other groups, particularly at the delayed posttest. This reduction indicated that there was less variation in scores within the group.

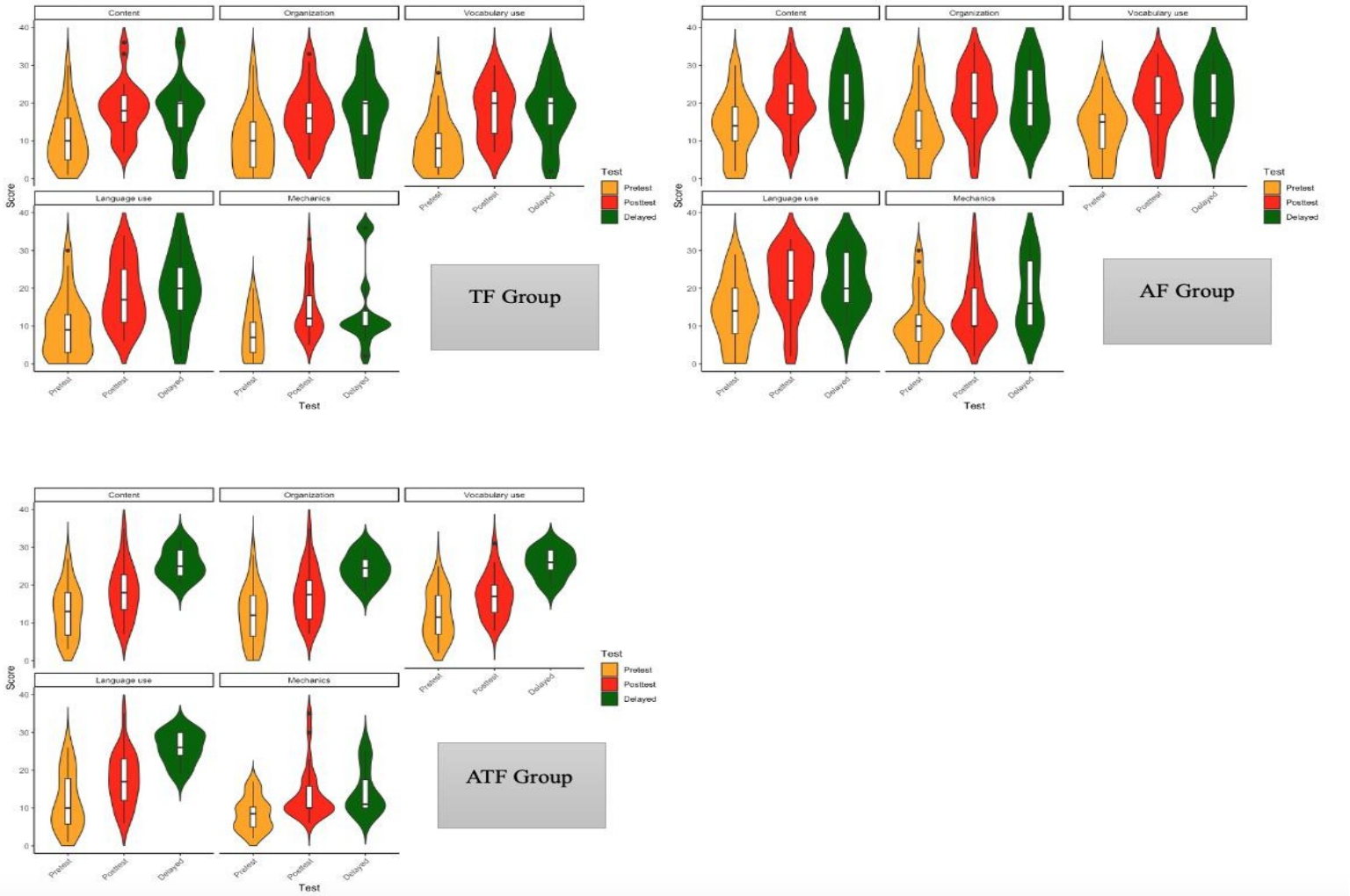


Figure 12 Performance of Feedback Groups in Terms of the Five Writing Components (Content, Organisation, Vocabulary Use, Language Use, Mechanics).

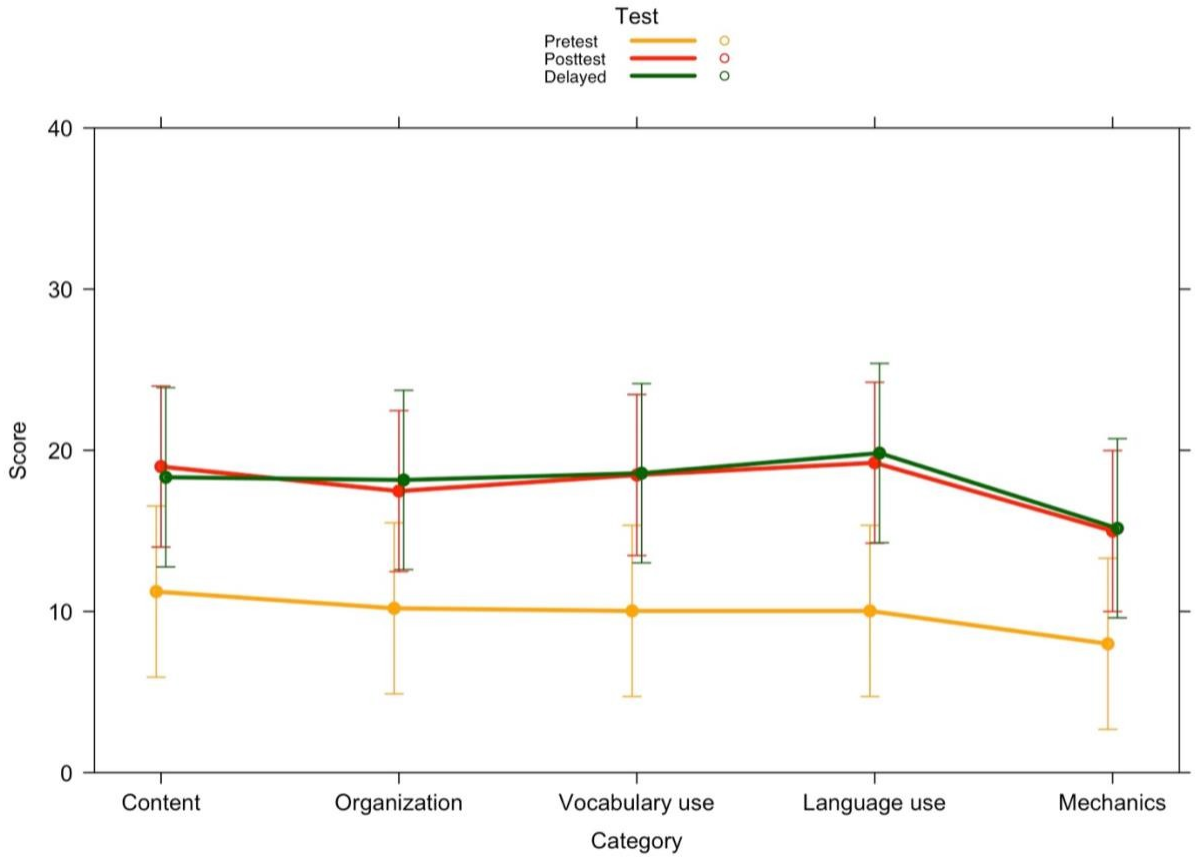


Figure 13 Performance of Teacher Feedback Group in Terms of the Five Writing Components.

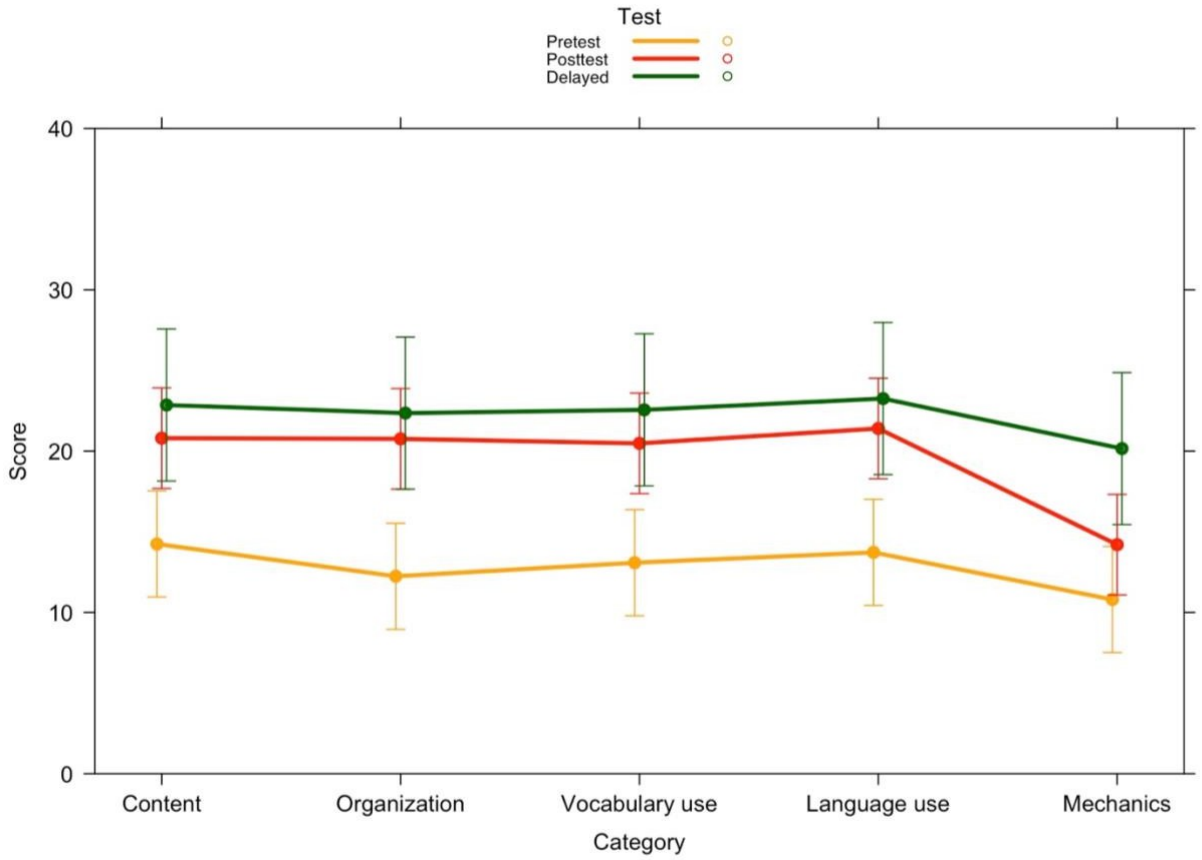


Figure 14 Performance of Automated Feedback Group in Terms of the Five Writing Components (Text Content, Text Organisation, Vocabulary Use, Language Use, and Mechanics).

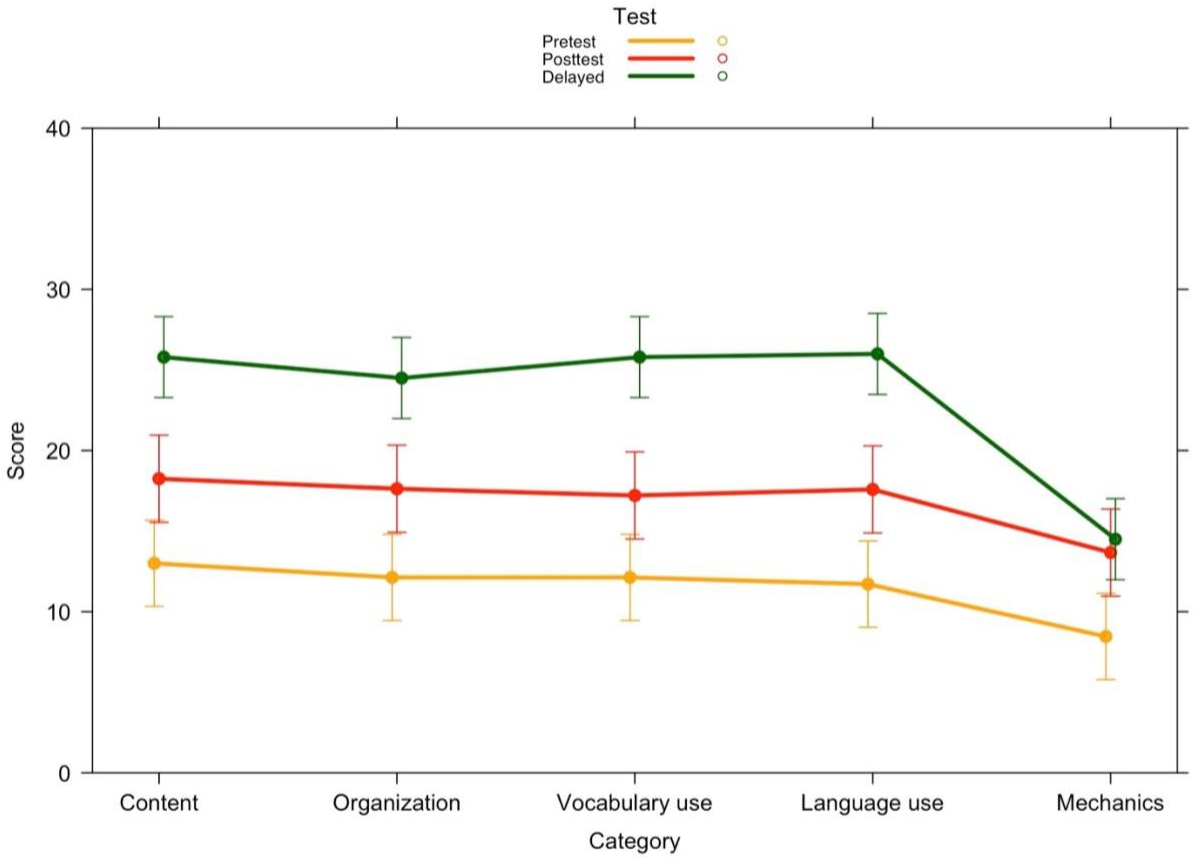


Figure 15 Performance of Hybrid Feedback Group in Terms of the Five Writing Components (Text Content, Text Organisation, Vocabulary Use, Language Use, and Mechanics).

5.4 Research Question 3: Comparison of the Effects of Feedback Condition and Writing Genre (Expository, Persuasive) on Overall L2 Writing Production.

This section sets out to answer the third research question: Do the effects of three different types of feedback on writing differ depending on the genre of writing?

The results of the mixed-effects model are presented for each group separately. The genre of writing was added to the model as a fixed effect, along with the test time, including the interaction between them. In reporting the results of the mixed-effects models, the focus is on the fixed effect estimates which directly relate to the research question by indicating whether or not there are any differences between the effects of feedback in the two writing genres (expository and persuasive) on overall L2 writing production.

Table 24 presents descriptive statistics for the writing scores regarding the expository genre over time. The scores for the teacher feedback group changed from the pretest (M = 29.6, SD = 22.4) to the immediate posttest (M = 42.4, SD = 19.6) and the delayed posttest (M = 43.2, SD = 22.2). The scores of the automated feedback group were (M = 35.8, SD = 20.4) at the pretest, (M = 50.3, SD = 19) at the immediate posttest and (M = 52.9, SD = 19.9) at the delayed posttest. The scores for the hybrid feedback group were (M = 32.9, SD = 16.6) at the pretest, (M = 42.2, SD = 18.9) at the immediate posttest and (M = 58.4, SD = 8.81) at the delayed posttest.

Table 24 *Summary of Descriptive Statistics for Writing Performance in the Expository Genre at Pretest, Immediate Posttest and Delayed Posttest.*

	Pretest				Immediate Posttest				Delayed Posttest			
Group	*M*	*SD*	*MD*	*IQR*	*M*	*SD*	*MD*	*IQR*	*M*	*SD*	*MD*	*IQR*
TF (*N* = 25)	29.6	22.4	27	42	42.4	19.6	45	30	43.2	22.2	45	14
AF (*N* = 25)	35.8	20.4	42	21	50.3	19	52	25	52.9	19.9	46.5	24
ATF *(N* = 24)	32.9	16.6	28.5	20.8	42.2	18.9	42.5	22.8	58.4	8.81	60	14

Table 25 provides the mean, standard deviation, median and interquartile range for total writing scores in the persuasive genre across the three feedback conditions. In the teacher feedback group, scores increased from (*M* = 18.1, *SD* = 19) to (*M* = 43, *SD* = 19.1) at the immediate posttest and (*M* = 43.5, *SD* = 22.5) at the delayed posttest. For the automated feedback group, scores rose from (*M* = 28.3, *SD* = 21.4) to (*M* = 47.3, *SD* = 20.1) at the immediate posttest and (*M* = 52, *SD* = 21.6) at the delayed posttest. Similarly, the hybrid feedback group exhibited an increase in scores from (*M* = 24.5, *SD* = 16.8) to (*M* = 42.1, *SD* = 17.6) at the immediate posttest and (*M* = 57.7, *SD* = 10) at the delayed posttest.

Table 25 *Summary of Descriptive Statistics for Writing Performance in the Persuasive Genre at Pretest, Immediate Posttest and Delayed Posttest for the Three Feedback Groups.*

	Pretest				Immediate Posttest				Delayed Posttest			
Group	*M*	*SD*	*MD*	*IQR*	*M*	*SD*	*MD*	*IQR*	*M*	*SD*	*MD*	*IQR*
TF (*N* = 25)	18.1	19	10	23	43	19.1	42	27	43.5	22.5	45	25
AF (*N* = 25)	28.3	21.4	28	31	47.3	20.1	50	23	52	21.6	47.5	38
ATF (*N* = 24)	24.5	16.8	24.5	30.2	42.1	17.6	41.5	18	57.7	10	60	20

5.4.1 Teacher Feedback Group

Mixed-effects model analysis indicated that the fixed effects explained a relatively small portion of the variance in total writing scores regarding the genre of writing (marginal R^2 = .185), whereas the combined fixed and random effects explained a larger amount (conditional R^2 = .886).

The model estimates in Table 26 indicate a significant change in expository writing scores within the teacher feedback group at the immediate posttest (β = 12.84 [5.18, 20.50], t = 3.29, p < .005) and at the delayed posttest (β = 16.39 [5.12, 27.67], t = 2.86, p = .011).

For persuasive writing, the model estimates indicate a significant increase at the immediate posttest (β = 24.80 [17.22, 32.54], t = 6.38, p < .0005) and at the delayed posttest (β = 26.92 [15.86, 37.97], t = 4.78, p < .0005) for the teacher feedback group.

Comparisons across writing genres indicate an initial difference in participants' writing skills at the pretest, with higher performance in expository writing compared to persuasive writing (β = -11.52 [-16.29, -6.75], t = 4.74, p < .0005). However, this discrepancy diminished

by the posttest (β = 0.52 [-4.25, 5.29], t = 0.21, p = .831) and the delayed posttest (β = -1.00 [-6.41, 4.42], t = -0.36, p = .718), suggesting that participants' persuasive writing skills improved over time.

Table 26 *Teacher Feedback Group: Results of the Linear Mixed-effects Model for the Writing Genres Examined (Expository, Persuasive).*

	Pretest	Immediate posttest	Delayed posttest
Expository genre	β = 30.24, 95% *CI* [19.39 – 41.09], *SE* = 5.52, t(3.23) = 5.47, p = .0098	β = 12.84, 95% *CI* [5.18 – 20.50], *SE* = 3.9, t(25) = 3.29, p < .005	β = 16.39, 95% *CI* [5.12 – 27.67], *SE* = 5.74, t(16) = 2.86, p = .011
Persuasive genre	β = 18.72, 95% *CI* [7.99 – 29.45], *SE* = 5.47, t(2.98) = 3.43, p = .042	β = 24.8, 95% *CI* [17.22 – 32.54], *SE* = 3.9, t(25) = 6.38, p < .0005	β = 26.92, 95% *CI* [15.86 – 37.97], *SE* = 5.63, t(16) = 4.78, p < .0005
Expository vs. persuasive	β = -11.52, 95% *CI* [-16.29 – -6.75], *SE* = 2.43, t(29) = -4.74, p < .0005	β = .52, 95% *CI* [-4.25 – 5.29], *SE* = 2.43, t(29) = .21, p = .831	β = -1.00, 95% *CI* [-6.41 – 4.42], *SE* = 2.76, t(47) = -.36, p = .719

Notes. Alpha values adjusted to correct for four repeated tests. *p < .0125. **p < .0025. ***p < .00025.

Full model estimates for teacher feedback group are provided in Appendix E.19.

5.4.2 Automated Feedback Group

Mixed-effects model analysis demonstrates that the fixed effects explained a relatively small portion of variance for the total writing scores regarding the genre of writing (marginal R^2 = .203), whereas the combined fixed and random effects explained a larger amount (conditional R^2 = .963).

The model estimates in Table 27 indicate a significant change within the automated feedback group in expository writing at both the immediate posttest (β = 14.52 [8.98, 20.06], t = 5.15, $p < .0005$) and the delayed posttest (β = 23.76 [12.63, 34.88], t = 4.19, $p < .005$).

In terms of persuasive writing, the model estimates indicate a significant increase at the immediate posttest (β = 19.04 [13.50, 24.58], t = 6.75, $p < .0005$) and at the delayed posttest (β = 32.79 [21.56, 44.02], t = 5.74, $p < .0005$) for the automated feedback group.

Similar to the teacher feedback group, between-genre comparisons for the automated feedback group revealed an initial difference between expository and persuasive writing at the pretest (β = -7.52 [-12.57, -2.47], t = -2.92, p = .007). However, this difference diminished over time, as the scores between the two genres were not significantly different at the immediate posttest (β = -3.00 [-8.06, 2.06], t = -1.17, p = .26) or at the delayed posttest (β = 1.48 [-3.86, 6.83], t = 0.55, p = .590).

Table 27 *Automated Feedback Group: Results of the Linear Mixed-effects Model for the Writing Genres Examined.*

	Pretest	Immediate posttest	Delayed posttest
Expository genre	β = 35.80, 95% *CI* [27.92 – 43.68], *SE* = 4.01, $t(24.16)$ = 8.92, $p < .0005$	β = 14.52, 95% *CI* [8.98 – 20.06], *SE* = 2.82, $t(25)$ = 5.15, $p < .0005$	β = 23.76, 95% *CI* [12.63 – 34.88], *SE* = 5.67, $t(8)$ = 4.19, $p < .005$
Persuasive genre	β = 28.28, 95% *CI* [19.86 – 36.70], *SE* = 4.29, $t(24.13)$ = 6.60, $p < .0005$	β = 19.04, 95% *CI* [13.50 – 24.58], *SE* = 2.82, $t(25)$ = 6.75, $p < .0005$	β = 32.79, 95% *CI* [21.56 – 44.02], *SE* = 5.72, $t(8)$ = 5.74, $p < .0005$
Expository vs persuasive	β = -7.52, 95% *CI* [-12.57 – -2.47], *SE* = 2.57, $t(25)$ = -2.92, p = .007	β = -3.00, 95% *CI* [-8.06 – 2.06], *SE* = 2.57, $t(25)$ = -1.17, p =.26	β = 1.48, 95% *CI* [-3.86 – 6.83], *SE* = 2.72, $t(32)$ = .55, p = .590

Notes. Alpha values adjusted to correct for two repeated tests. *$p < .025$. **$p < .005$. ***$p < .0005$.
Full model estimates for automated feedback group are provided in Appendix E.20.

5.4.3 Hybrid (Automated+teacher) Feedback Group

Mixed-effects model analysis demonstrates that the fixed effects explained a relatively small portion of the variance for the total writing scores regarding the genre of writing (marginal R^2 = .303), whereas the combined fixed and random effects explained a larger amount (conditional R^2 = .903).

The model estimates in Table 28 indicate a significant change within the hybrid feedback group in expository writing at both the immediate posttest (β = 9.37 [1.81, 16.94], t = 2.43, p = .023) and the delayed posttest (β = 24.25 [17.08, 31.42], t = 6.64, $p < .0005$).

For persuasive writing, the model estimates indicate a significant increase at the immediate posttest (β = 17.54 [9.97, 25.11], t = 4.55, p < .0005) and at the delayed posttest (β = 34.96 [27.78, 42.13], t = 9.57, p < .0005) for the hybrid feedback group.

Consistent with the other feedback groups, between-genre comparisons for the hybrid feedback group revealed an initial difference between expository and persuasive writing at the pretest phase (β = -8.33 [-13.03, -3.64], t = -3.49, p = .0017). However, this difference diminished over time, as the scores between the two genres were not significantly different at the immediate posttest (β = -0.17 [-4.86, 4.53], t = -0.07, p = .944) or at the delayed posttest (β = 2.37 [-2.83, 7.58], t = 0.90, p = .376).

Table 28 *Hybrid feedback group: results of linear mixed-effects model for the writing genre examined.*

	Pretest	Immediate posttest	Delayed posttest
Expository genre	β = 32.87, 95% *CI* [26.23 – 39.52], *SE* = 3.38, t(23.37) = 9.72, p < .0005	β = 9.38, 95% *CI* [1.81 – 16.94], *SE* = 3.85, t(24) = 2.43, p = .023	β = 24.25, 95% *CI* [17.08 – 31.42], *SE* = 3.65, t(15) = 6.64, p = .005
Persuasive genre	β = 24.54, 95% *CI* [18.00 – 31.09], *SE* = 3.33, t(23) = 7.37, p < .0005	β = 17.54, 95% *CI* [9.97 – 25.11], *SE* = 3.85, t(24) = 4.55, p < .0005	β = 34.96, 95% *CI* [27.78 – 42.13], *SE* = 3.65, t(15) = 9.57, p < .0005
Expository vs persuasive	β = -8.33, 95% *CI* [-13.03 – -3.64], *SE* = 2.39, t(26.44) = -3.49, p = .0017	β = -0.17, 95% *CI* [-4.86 – 4.53], *SE* = 2.39, t(26) = -0.07, p = .944	β = 2.37, 95% *CI* [-2.83 – 7.58], *SE* = 2.65, t(39) = 0.90, p = .377

Notes. Alpha values adjusted to correct for two repeated tests. *p < .025. **p < .005. ***p < .0005.
Full model estimates for hybrid feedback group are provided in Appendix E.21.

In summary, incorporating genre as an independent variable when comparing the three feedback conditions (teacher, automated, and hybrid) revealed no significant within-group differences in writing scores related to genre. This indicates that all groups showed significant improvement in both expository and persuasive writing at the immediate and delayed posttests. However, initial scores at the pretest stage showed that participants in all three groups performed better in the expository genre, suggesting relatively greater gains in the persuasive genre at the later testing phases. Between-genre analyses further revealed a significant difference in students' performance across genres at the pretest stage. This difference, however, disappeared at both the immediate and delayed posttests, as no significant differences were found between the genres for any of the groups (see Figure 17).

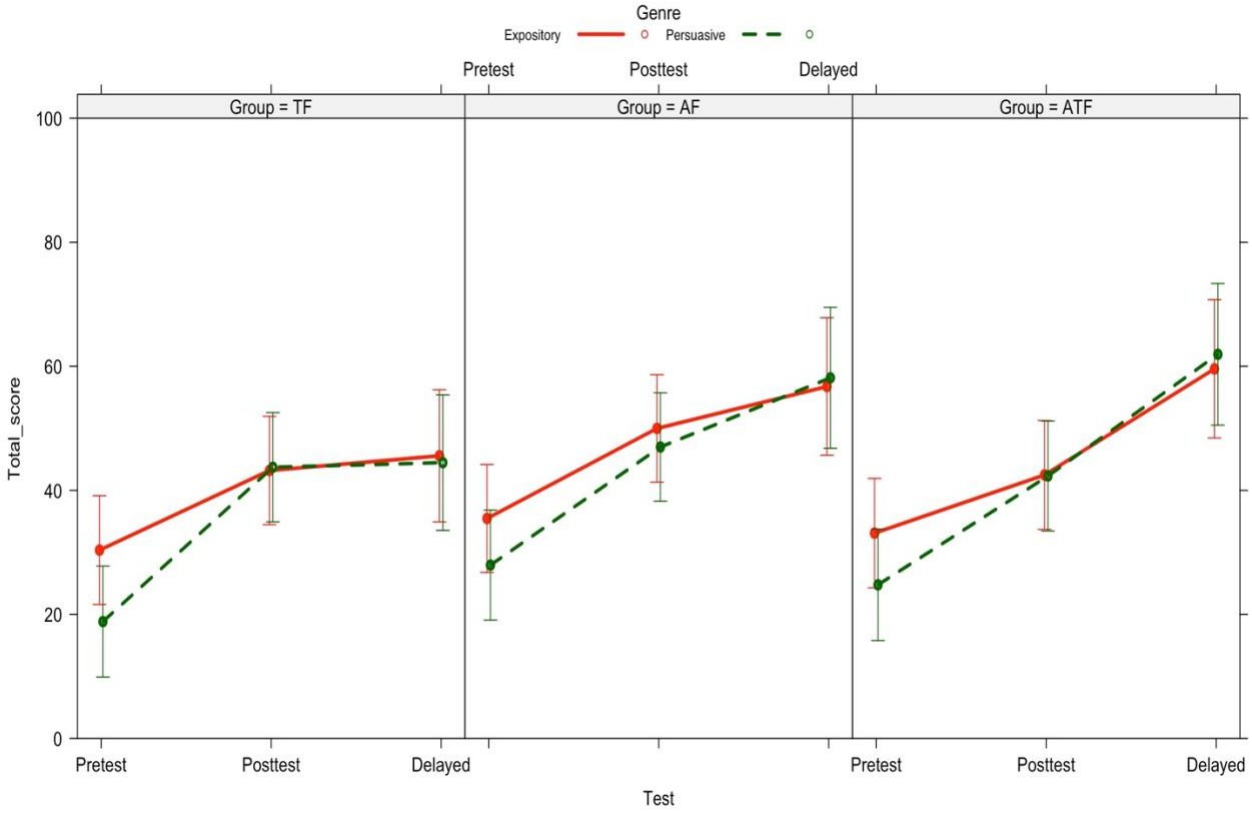


Figure 16 Plot of the Linear Mixed-effects Model Total Writing Scores in the Two Writing Genres Tested (Expository, Persuasive).

5.5 Summary of the Results

The current study looked at the effect of various feedback conditions (teacher feedback, automated feedback and hybrid feedback) on the writing performance of EFL learners. The participants' performance was examined at three test times (pretest, immediate posttest and delayed posttest). The first research question focused on overall writing performance and its potential correlation with feedback conditions. The second research question delved into participants' performance across five writing components (text content, text organisation, vocabulary use, language use and mechanics), investigating whether any of the three feedback conditions had a differential impact on a specific writing component. The third research question extended the inquiry to changes in writing genre, exploring whether such changes influenced participants' writing under the three feedback conditions. The analysis utilised descriptive statistics, followed by the results of linear mixed-effects models and Cohen's *d* effect sizes.

Regarding the first research question, the impact of three feedback conditions (teacher-only, automated-only, and hybrid) on overall L2 writing performance was examined. Descriptive statistics indicated that writing scores improved over time within all groups. A linear mixed-effects model analysis showed that both fixed and random effects significantly contributed to explaining the variance in total writing scores.

Model estimates revealed that all three feedback groups demonstrated significant gains in overall writing performance at both the immediate and delayed posttests. Between-group comparisons indicated no statistically significant differences at the pretest or immediate posttest. However, at the delayed posttest, significant differences were observed between the teacher feedback group and the automated feedback group, as well as between the teacher feedback group and the hybrid feedback group, with the latter showing a larger effect size.

The results addressing the second research question indicated positive changes across all writing components within each of the three feedback groups. However, at the delayed posttest, the hybrid feedback group consistently demonstrated the largest effect across the writing components.

Between-group comparisons revealed no statistically significant differences among the groups at the pretest or immediate posttest across any of the writing components. At the delayed posttest, however, significant differences emerged between the teacher feedback group and the hybrid feedback group in the areas of text content, text organization, vocabulary use, and language use.

For the third research question, descriptive statistics indicated consistent increases in writing scores over time across all groups in both writing genres (expository, persuasive).

Linear mixed-effects model analysis revealed that participants in all three groups demonstrated significant improvements in both expository and persuasive writing at the immediate and delayed posttests. However, between-genre comparisons revealed a statistically significant difference in writing performance across the three feedback groups at the pretest, with participants performing better in expository writing than in persuasive writing. Model estimates showed that this initial disparity between genres decreased at both the immediate and delayed posttests for the three feedback groups. This reduction suggests that participants made greater improvements in persuasive writing, allowing their performance in that genre to catch up with, or in some cases exceed, their performance in expository writing over time.

6. Discussion

6.1 Overview

This study sought to compare the impact of teacher written feedback, automated feedback, and hybrid feedback on EFL students' writing quality. The first research question addressed the overall writing performance by looking at how the feedback condition may affect the total writing score. The second research question examined how the type of feedback affected specific discourse components of the written text—specifically, text content, text organization, vocabulary use, language use, and mechanics—in relation to the feedback provided. While the third research question examined whether the impact of feedback differed across the three feedback conditions depending on the change in genre of writing (expository or persuasive).

The chapter begins with a summary of the key findings, followed by a discussion of how different feedback conditions have impacted EFL learners in Saudi Arabia in light of these results. First, I will discuss the effects of the three feedback conditions—automated feedback, teacher feedback, and hybrid feedback—on students' overall writing performance. Next, I will examine the impact of feedback condition on specific writing components. Finally, I will explore if changes in writing genre would influence the effectiveness of feedback conditions on students' overall writing performance.

6.2 Key Findings

The findings revealed that all three feedback conditions—teacher feedback, automated feedback, and hybrid feedback—had a significant positive impact on participants' overall writing scores at both the immediate posttest and at the delayed posttest. However, between-group comparisons at the delayed posttest indicated significant differences between the

teacher feedback group and the automated feedback group, as well as between the teacher feedback group and the hybrid feedback group. These findings indicated that both automated feedback group and hybrid feedback group outperformed the teacher feedback group at the delayed posttest time.

The findings also showed that, across the five assessed components of writing—text content, text organization, vocabulary use, language use, and mechanics—all three groups demonstrated significant improvement at both the immediate and delayed posttests.

While between-group comparisons were largely non-significant for these writing components, the hybrid feedback group outperformed the other groups at the delayed posttest in text content, text organization, vocabulary use, and language use scores.

With regard to the interaction between genre (expository vs. persuasive) and feedback condition on students' writing performance, all three feedback groups demonstrated significant improvement in both genres at the immediate and delayed posttests. However, comparisons between genres at the pretest revealed significant differences in performance, with participants performing better on the expository task. Specifically, the teacher feedback group showed a significant gap favoring expository writing as did the automated feedback group and the hybrid feedback group.

However, the initial performance gap between genres diminished at both the immediate and delayed posttest phases across all groups, suggesting that participants performed better on the persuasive writing task over time.

6.3 The Impact of Feedback Type on L2 Students Writing Performance

The results indicated that all three feedback conditions led to significant improvements over time. The linear mixed-effects model estimates showed a significant positive change within all three groups at both the immediate and delayed posttests. These results suggest that feedback, regardless of type, has a beneficial impact over time on writing quality and can result in immediate improvement in students' total writing scores. The immediate improvement across groups reinforces the established role of feedback in fostering rapid advancements in student performance and aligns with other studies examining the impact of written feedback in language learning contexts. (Hyland& Hyland, 2006 a, b; Yu, 2021)

The findings also revealed that all three feedback groups showed significant improvements in writing scores across the examined writing components—text content, text organization, vocabulary use, language use, and mechanics—at the immediate posttest and the delayed posttest, confirming that feedback, regardless of type, has a positive impact on L2 writing performance. By the delayed posttest, the hybrid feedback group had retained the highest level of improvement, as reflected in both mixed-effect model estimates and effect sizes across all evaluated writing elements.

While the change in total scores was significant across all feedback conditions, it is worth noting that the teacher feedback group achieved the highest gain at the immediate posttest (β = 37.72) with the largest effect size (d = 1.1) compared to (β = 33.56, d = .90) for the automated feedback group and (β = 26.92, d = .90) for the hybrid feedback group.

One possible explanation for this outcome may be related to students' engagement with the feedback provided. Zhang and Hyland (2018) found that students' behavioral engagement with

feedback—such as actively reading, responding to, and implementing suggestions—played a critical role in the success of the writing process and in improving overall performance.

In addition to learner engagement, this finding can also be interpreted through the lens of sociocultural theory (SCT) and the zone of proximal development (ZPD). According to Vygotsky (1978), learning occurs through socially mediated interaction, particularly with more knowledgeable individuals such as teachers or peers (Lantolf & Thorne, 2006; Bornstein & Bruner, 2014). Within this framework, feedback functions as a mediational tool that facilitates language development by helping learners notice gaps in their performance and guiding them toward improvement. Ohta (2001) emphasized that teacher support—when aligned with the learner's ZPD—can significantly contribute to language improvement.

In this context, the effectiveness of teacher feedback may be attributed to its timing, clarity, and relevance to each learner's developmental level. As Aljaafreh and Lantolf (1994) suggested, feedback that is tailored to the learner's specific needs within their ZPD can provide appropriate scaffolding, allowing them to progress from what they can do independently to what they can accomplish with guidance. The teacher feedback group in this study likely benefited from this kind of individualized, context-sensitive support, which not only supported engagement but also provided scaffolding that was cognitively and linguistically appropriate. From this perspective, teacher feedback served as a strategic intervention that supported learners in bridging the gap between their current writing abilities and their potential performance.

In contrast, AWE can become more of a burden than a facilitator if students feel overwhelmed by the quantity and complexity of feedback, making efficient revisions difficult Hyland and Hyland (2006a). Moreover, some students may not fully engaged with automated feedback in the revision process (Stevenson, 2016). Ranalli (2018) observed that some

participants in his study chose to delete incorrect sentence fragments rather than actively revising their work, suggesting a tendency to avoid complex revisions rather than engaging with feedback to improve accuracy. In addition, when students receive comprehensive corrections from an AWE system, they may prioritize refining content, meaning, and idea development over addressing mechanical issues. This perspective is supported by the outcomes of the hybrid feedback group, where students demonstrated improvements across all examined writing components. The integration of automated and teacher feedback appeared to encourage students to engage more deeply with the feedback from both sources, ensuring a balanced focus on content and linguistic accuracy.

However, at the delayed posttest, between-group comparisons revealed a significant difference, with both the automated and hybrid feedback groups outperforming the teacher feedback group. While these findings suggest that computer-generated feedback may support greater retention of writing-related knowledge over time—particularly when compared to teacher-only feedback—it is important to acknowledge that not all participants completed the delayed posttest. This limitation may have influenced the reliability and generalizability of the between-group comparisons at that stage.

Despite this limitation, the observed trend still highlights the potential long-term benefits of automated and hybrid feedback approaches in L2 writing development. One possible explanation for this finding is offered by Ranalli (2018), who argued that automated feedback provides immediate and specific feedback, which learners tend to value more than delayed feedback. Immediate feedback reduces cognitive load, allowing students to process and apply corrections while their errors are still fresh in memory. This advantage is thought to contribute to

stronger retention of linguistic knowledge, as students are able to recognize, correct, and internalize mistakes more effectively.

This explanation was also echoed in Kleij (2012) and Zhu et al. (2020), where students reported higher motivation and engagement when receiving instant, specific feedback. Immediate feedback creates a continuous feedback loop, reinforcing learning by encouraging students to revise their work iteratively rather than waiting for delayed corrections. In contrast, teacher feedback—often delayed due to workload constraints—may not be as immediately accessible, leading to less effective knowledge retention over time.

It is important to note, however, that in the current study, the timing of feedback was carefully controlled to minimize such discrepancies. All three groups were required to take a 30-minute break before revising their work based on the feedback received. During this time, teachers provided written feedback to approximately 7-10 students per group (see Section 4.5.4.2), ensuring that feedback across conditions was administered within a similar timeframe.

The observed greater impact of the hybrid feedback condition may also tentatively relate to participants' increased independence when interacting with computer-generated feedback. Engaging directly with automated systems might have encouraged learners to be more responsible for their revisions, which in turn increased their autonomy and sustained their focus during the revision process, and thus elevated continuous noticing (Karatay & Karatay, 2024; Shannon & Chapelle, 2017). This idea was demonstrated as well in Warshauer (1997, 1999) who argued that interacting with technology transforms learning experience and creates a form of discourse that offers a more flexible and inclusive communication environment. In contrast, participants in the teacher feedback-only group did not have the opportunity to interact independently with feedback or to choose which aspects to focus on based on their own

engagement. Instead, they tended to passively accept the feedback provided by the teacher and revised their written output accordingly.

Building on these findings, the results may tentatively suggest that the hybrid feedback condition provided participants with a balance between immediate correction and personalized guidance. This approach enabled students to receive both the prompt, detailed feedback characteristic of automated systems and the more individualized, nuanced input typically offered by teachers. The combination of these feedback sources may have contributed to a more comprehensive and supportive feedback experience, potentially facilitating sustained improvement in writing performance.

The observed benefits of hybrid feedback align with previous literature advocating for blended feedback approaches in language learning (Grimes & Warschauer, 2010; Link et al., 2020; Palermo & Thompson, 2018; Sari & Han, 2024; Ware, 2011; Weigle, 2013; Wilson & Czik, 2016). These studies suggest that combining automated with human feedback can maximize the advantages of both methods.

Grimes and Warschauer (2010), for example, recommended the thoughtful integration of automated writing evaluation (AWE) and teacher-provided feedback, arguing that such integration can motivate students to write and revise more frequently, increase overall writing practice, and enable teachers to focus their attention on higher-level concerns rather than merely on mechanical issues. Similarly, Link et al. (2020) found that blending AWE with teacher feedback allowed teachers to concentrate their feedback more effectively on content and idea development. Their findings further indicated that students who had access to AWE maintained language accuracy over time compared to those who relied exclusively on teacher feedback.

Palermo and Thompson (2018) likewise reported that students receiving hybrid feedback produced higher-quality essays compared to those receiving feedback from teacher only. Along the same line, Sari and Han (2024) found that combining automated and teacher feedback led to greater improvements in student writing. Notably, participants in their study explicitly confirmed their preference for a combined feedback approach, highlighting the perceived benefits of receiving both automated and personalized teacher feedback.

Karataly and Karataly (2024), who emphasized the importance of integrating AWE tools with teacher feedback and highlight the need for a balanced approach to using both types of feedback to effectively support L2 writing instruction. Their findings further reinforce the idea that a hybrid feedback model is thought to foster deeper engagement with the revision process, making it a valuable strategy for improving writing proficiency.

Furthermore, the lack of long-term efficacy of teacher-only or automated-only feedback may highlight the limitations of each approach when used in isolation. Teacher feedback, while often detailed and context specific (Dikli & Bleyle 2014) may be less frequent or timely due to teacher workload constraints (Chandler 2003). In contrast, automated feedback, though immediate and comprehensive (Dikli & Bleyle 2014), can overwhelm students with an excessive number of suggestions, making it difficult for them to prioritize revisions effectively (Hyland & Hyland 2006a; Dikli 2010; Ranalli 2018). Additionally, automated feedback lacks the individualization and contextual awareness necessary to provide meaningful feedback on students' ideas and text cohesion. Unlike teachers, AWE systems are limited in their ability to evaluate rhetorical effectiveness, argument structure, and overall coherence, which are essential aspects of writing development (Wilson & Czik 2016).

To summarize, feedback regardless of type enabled EFL students to improve their writing performance. Within-group comparisons at immediate posttest time suggest that written feedback contributes to short-term writing progress (Mackey, 2012). Additionally, computer-generated feedback can be considered as effective as teacher-written feedback, as its iterative process of drafting, receiving feedback, and redrafting leads to continuous improvement in written output (Mackey, 2012; Swain, 1985, 1995). This improvement aligns well with Schmidt's (2001) noticing hypothesis, suggesting that feedback on writing helps learners notice discrepancies between their current interlanguage and the target language forms. Although Schmidt originally focused on oral feedback, noticing in oral production contexts can reasonably be extended to writing, where linguistic discrepancies may be even more readily apparent in written form. Additionally, the extended time available for students to process, revise, and respond to written feedback may reduce cognitive load, thus developing learners' ability to internalize corrections—a benefit particularly relevant in L2 writing contexts. Thus, distributing feedback between AWE and the teacher may have mitigated the potential for cognitive overload, as students were not overwhelmed by an excessive quantity of corrections. The division of feedback responsibilities allowed learners to process and apply feedback more effectively, reinforcing the pedagogical advantage of a combined feedback approach.

6.4 The Impact of Genre Type and Feedback Condition on L2 Writing Performance

The third research question examined the effect of genres (expository and persuasive) and feedback conditions (teacher only, automated only, and hybrid feedback) on students' writing quality. To investigate this, the study analyzed students' final written outcome scores at three different time points: pretest, immediate posttest, and delayed posttest. The scores were evaluated within each genre and compared across the different feedback groups to determine the impact of feedback type and genre on writing performance over time.

The results indicate that all three feedback groups—teacher-only, automated-only, and hybrid—demonstrated significant improvement in both expository and persuasive writing tasks at both the immediate and delayed posttests. Although participants initially performed better on the expository genre, mixed-effects model estimates revealed that greater gains were consistently observed in the persuasive genre across all feedback conditions. This suggests that genre complexity may have influenced the degree of improvement, with persuasive writing eliciting more substantial development over time.

At the pretest stage, a significant difference between genres was evident in all three groups, with higher performance on the expository task: teacher feedback (β = -11.52), automated feedback (β = -7.52), and hybrid feedback (β = -8.33). These differences, however, diminished at the immediate and delayed posttests, indicating that participants made more marked progress in the persuasive genre. This trend aligns with Robinson’s cognition hypothesis (2001, 2007), which posits that more cognitively demanding tasks promote deeper engagement and richer language output. The persuasive genre, requiring learners to construct arguments, integrate evidence, and anticipate counterarguments, may have encouraged more sustained cognitive and linguistic effort, which in turn amplified the impact of feedback.

Furthermore, the word count difference between genres—250 words for persuasive and 150 for expository—may have contributed to this pattern by affording students more space for elaboration, reflection, and revision. These additional opportunities for engagement may have led to more effective uptake of feedback, especially in tasks that required critical thinking and rhetorical structuring.

This finding also reflects trends observed in prior research, where more complex writing tasks (e.g., argumentative or narrative writing) are commonly used to investigate the effectiveness of feedback, due to their greater potential for generating revisions and linguistic development (Barrot, 2023; Palermo & Thompson, 2018; Ranalli, 2018; Zhu et al., 2020). The results of the present study reinforce this rationale, suggesting that genre not only mediates learners' engagement with feedback but may also amplify the instructional value of different feedback types.

In summary, while all feedback types supported writing development across genres, the interaction between task complexity (as represented by genre) and feedback condition appears to play a meaningful role in determining the extent of learners' progress. These findings support the integration of cognitively demanding writing tasks in feedback-focused instruction, especially when paired with multimodal feedback approaches such as hybrid systems.

6.7 Summary

Many studies have emphasized the importance of integrating automated feedback with traditional teacher feedback to maximize the benefits of automated writing evaluation (AWE), while allowing teachers to focus on more complex, context-specific issues that AWE systems alone may not adequately address (e.g., Grimes & Warschauer, 2010; Link et al., 2020; Wilson & Czik, 2016). However, the optimal approach to implementing AWE in language instruction remains inconclusive.

Building on this gap, the present study examined the impact of a hybrid feedback approach (automated + teacher) by comparing it with teacher-only and automated-only feedback. The findings confirmed the effectiveness of the hybrid feedback mode in developing EFL students' writing performance. In addition, the study explored how each feedback condition influenced specific aspects of writing, including text content, text organization, vocabulary use, language use, and mechanics.

Another key dimension of this study was the examination of writing genres to determine whether genre differences influenced the effectiveness of the feedback provided. While within-group comparisons revealed significant improvement in both genres across all groups over time, between-genre comparisons suggested that genre type may affect students' revision behaviors and engagement with feedback. Specifically, the more cognitively demanding genre—persuasive writing—resulted in greater gains in mean scores across groups, aligning with Robinson's cognition hypothesis, which posits that increased task complexity can lead to better language production.

These findings contribute to the ongoing discussion on the implementation of AWE, reinforcing the value of a combined feedback approach and highlighting the influence of genre in shaping students' interaction with feedback.

7. Conclusion

7.1 Overview

This study aimed to explore the impact of automated written feedback on the quality of EFL learners' essay writing in Saudi Arabia. In particular, the current research examined how EFL learners deploy various modes of written feedback (teacher only, automated only, hybrid (automated+teacher) feedback) to improve their writing outcome. Additionally, the study investigated how the three feedback modes influenced specific writing elements, including text content, text organization, vocabulary use, language use, and mechanics. The study also explored whether changes in writing genres: (expository and persuasive) affected the participants' writing performance relative to their feedback condition.

A quasi-experimental design was chosen to compare the effects of the three feedback conditions. Evaluation of writing quality, tested at three stages (pretest, immediate posttest, and delayed posttest), involved five variables: text content, text organization, vocabulary use, language use, and mechanics. Participants were 74 foundation-year students of pre-intermediate English language proficiency. During the treatment, participants were randomly divided into three groups, each receiving one of the three feedback types on their writing. The treatment lasted for nine weeks. During the treatment sessions, participants practiced writing essays and then received feedback according to their group, followed by revision and rewriting of the essays. The writing test scores were analyzed to explore the impact of each feedback type on learners' writing competency.

This chapter presents a summary of the main findings, followed by the theoretical and pedagogical implications. Then, a discussion of the study's limitations is outlined, along with suggestions for future research.

7.2 Summary of Main Findings

The present study was guided by three research questions. Overall, feedback—regardless of type—had a significant positive impact on participants' writing performance over time. The following summarises the main findings of the study.

RQ 1 focused on exploring the impact of three feedback conditions—teacher only, automated only, and hybrid feedback—on the overall writing quality. Analysis of data revealed the following results:

- The teacher feedback condition had a significant impact on students' overall writing performance at both the immediate and delayed posttests, with the largest mean gain score among the groups at the immediate posttest.
- Automated feedback condition also demonstrated a significant positive impact on participants' overall writing outcome at both the immediate and delayed posttests.
- Similarly, the hybrid feedback condition demonstrated a significant positive impact on participants' overall writing outcomes at both time points, with the largest mean gain score at the delayed posttest.
- No significant differences were found between the three groups at the immediate posttest.
- At the delayed posttest, both the automated-only and hybrid feedback groups significantly outperformed the teacher feedback group; however, no significant difference was observed between the two intervention groups AF and ATF.

RQ 2 explored the impact of three feedback conditions—teacher only, automated only, and hybrid feedback—on text content, text organization, vocabulary use, language use, and mechanics. The main findings which the study obtained are as follows:

- The teacher feedback group showed significant improvement from the pretest to both the immediate and delayed posttests across all writing components: text content, text organization, vocabulary use, language use, and mechanics.
- The automated feedback group also demonstrated significant gains from the pretest to both posttests in all measured areas.
- The hybrid feedback group exhibited significant improvement from the pretest to the immediate and delayed posttests across all five writing elements.
- Between-group comparisons revealed no significant differences between the groups at the immediate posttest. However, at the delayed posttest, a significant difference was observed only between the teacher feedback group and the hybrid feedback group in relation to text content, vocabulary, and language use scores. All other comparisons showed no significant differences.

RQ 3 investigated the effects of writing genre (expository, persuasive) and feedback condition on overall writing production. The study yielded the following results:

- All feedback groups demonstrated significant improvement in both genres at the immediate and delayed posttest.
- Between-genre comparisons revealed no significant differences between genres at both the immediate and delayed posttests across all groups. However, a significant difference between genres was observed at the pretest stage, which diminished by the immediate posttest phase.

7.3 Study Contributions

This study has made methodological, theoretical and pedagogical contributions.

Methodologically, this study contributes to the field by advancing our understanding of the role and potential of AWE systems in EFL writing classrooms. It offers empirical insights into the effects of automated feedback on writing quality and systematically compares its effectiveness to traditional teacher feedback. While prior research has suggested that AWE works best as a complement to teacher input, this study builds on that foundation by proposing and evaluating a blended feedback model that integrates both feedback types. Additionally, the research introduces a novel methodological angle by examining how writing genre interacts with feedback type—an area that has received limited attention. This integrated approach provides a broader and more nuanced perspective on how different feedback modes function across varying writing contexts, offering valuable direction for future research designs and pedagogical applications.

Theoretically, the findings of this study contribute to several key frameworks related to second language learning. First, the results offer empirical support for the interaction hypothesis (Long, 1996), affirming that interaction—particularly through feedback—plays a vital role in promoting language development.

The study further underscores the importance of technology as a medium for interaction, highlighting how automated feedback systems can facilitate meaningful engagement with language input and promote revision-focused learning.

Additionally, the findings align with sociocultural theory and the concept of the zone of proximal development (ZPD). Learners who received feedback that slightly exceeded their current proficiency level were able to process and apply it effectively, leading to measurable

improvements in their writing. This suggests that both automated and teacher feedback can function as scaffolding tools that help students move from assisted to independent performance.

Moreover, by examining performance across two distinct genres—expository and persuasive—the study engages with the cognition hypothesis (Robinson, 2001, 2007), which posits that increasing task complexity can promote deeper cognitive processing and language development. Genre variation introduced different cognitive demands, and the data revealed how learners responded to these complexities under different feedback conditions. This theoretical insight into how cognitive and interactionist perspectives intersect in writing development.

Pedagogically, the current research findings have important implications for EFL writing instruction. It provided empirical evidence for the potential of using automated feedback in developing writing strategies. The apparent short-term effectiveness of all feedback types suggests that any structured feedback can positively impact immediate student performance. The longer-term results highlight the potential value of the hybrid feedback in supporting sustained writing improvement. For EFL instructors, integrating automated feedback tools or supplementing teacher feedback with automated options may be a more effective strategy for achieving lasting improvements in writing quality, especially if logistical or time constraints limit the frequency of traditional teacher feedback.

Moreover, for programs or instructors that currently rely exclusively on teacher feedback, this study's findings suggest that incorporating automated elements ay tentatively increase the durability of learning gains. Since EFL writing often involves learning complex structures and rules, which may benefit from repeated practice and correction, automated systems could provide supplementary reinforcement that might be less feasible to deliver consistently through teacher feedback alone. Introducing students to such systems early in the writing process and

encouraging ongoing engagement with automated feedback may help them monitor progress, internalize feedback patterns, and ultimately build greater independence in revision. This approach aligns with previous research supporting the use of automated writing evaluation (AWE) to facilitate iterative drafting and revision, and positions technology as a valuable pedagogical tool in EFL writing instruction.

Based on the pedagogical implications of this study, several practical recommendations are proposed to guide EFL instructors and curriculum designers in effectively integrating feedback strategies, particularly automated writing evaluation (AWE) systems, into writing instruction.

1. **Integrate AWE systems early in instruction.** Students should be introduced to automated feedback tools at the beginning of a writing course. Early familiarization helps learners become comfortable with the tool's interface and feedback categories, promoting consistent and confident use throughout the course.
2. **Adopt a hybrid feedback approach**. A combination of teacher and automated feedback may yield the most effective results. While AWE tools can efficiently address lower-order concerns such as grammar, mechanics, and lexical accuracy, teacher feedback should target higher-order elements including content development, organization, and coherence. This complementary approach maximizes the strengths of both feedback types.
3. **Use AWE for repeated practice and reinforcement**. AWE systems can be leveraged to provide frequent, low-stakes feedback between formal teacher evaluations. Assigning revision tasks that utilize AWE encourages learners to engage with their writing more frequently and supports the development of self-editing habits.

4. **Set clear guidelines and scaffolds for feedback use**. Instructors should offer explicit instruction on how to interpret and apply automated feedback. This can include setting revision goals, providing feedback checklists, or requiring a minimum feedback score to promote deeper engagement with the system and its suggestions.
5. **Encourage reflective feedback engagement**. To promote metacognitive awareness, students can be asked to reflect on the feedback they receive. This may take the form of brief revision logs or reflection journals, encouraging learners to think critically about the changes they make and the rationale behind them.
6. **Align feedback strategies with genre and task complexity**. Feedback approaches should be tailored to the demands of different writing tasks. For genres that require higher cognitive effort, such as persuasive writing, hybrid feedback may offer the optimal level of scaffolding to support learner success while fostering independent development.
7. **Provide teacher training for effective AWE integration**. Professional development opportunities should be offered to support instructors in incorporating AWE into their instructional practices. Training should focus on understanding AWE's capabilities and limitations, aligning feedback strategies with learning goals, and effectively balancing teacher and automated input.

These recommendations are intended to guide the practical application of this study's findings and support the ongoing development of effective, scalable, and pedagogically sound feedback practices in EFL writing instruction.

Figure 15 presents a proposed model for integrating AWE into EFL writing instruction. Implementing this model in the classroom may support teachers in offering a broader range of writing tasks and providing students with more opportunities for practice, which could potentially contribute to improvements in overall writing competence.

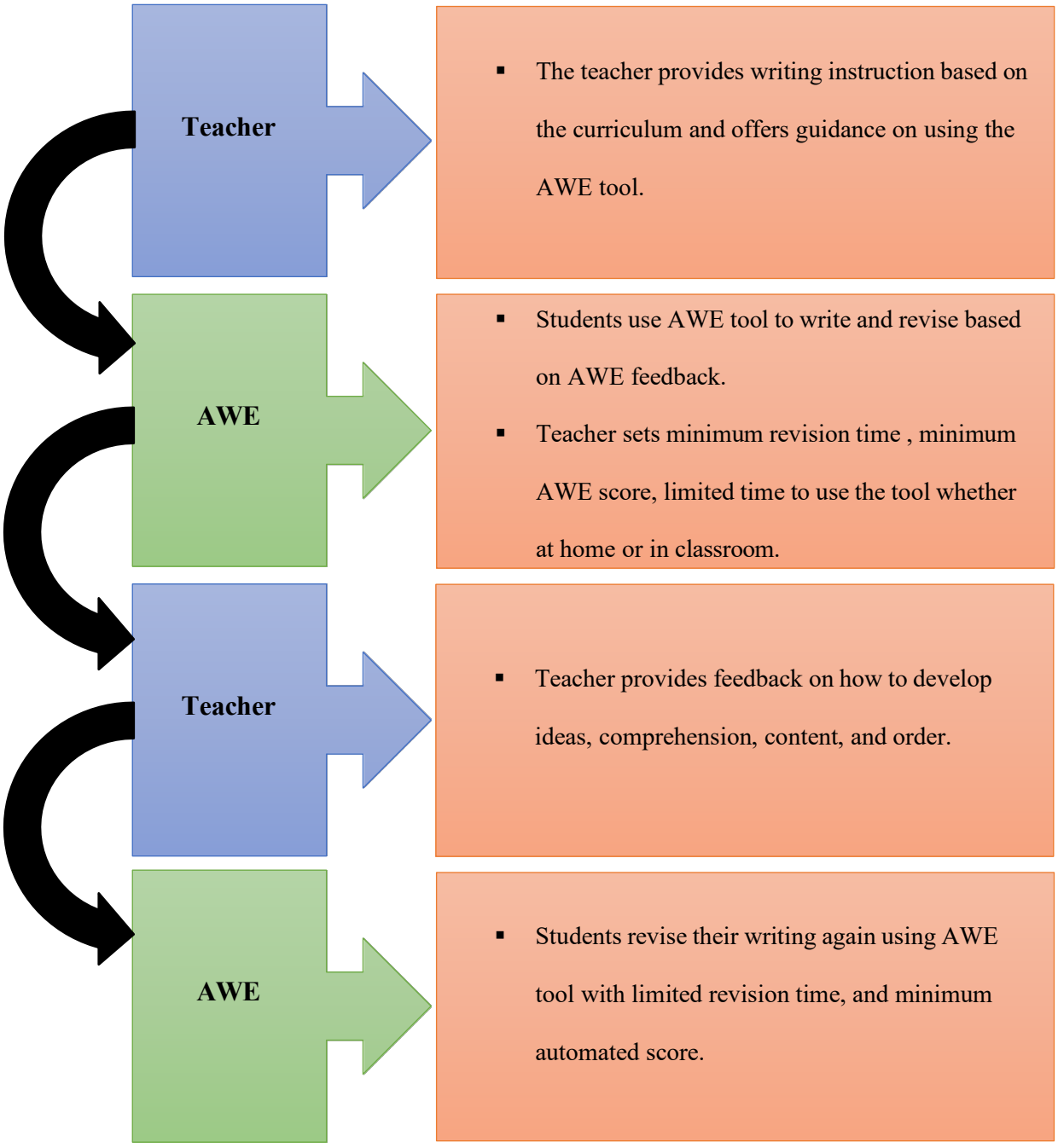


Figure 17 A Proposed Plan for Integrating AWE into EFL Writing Instruction.

7.3 Limitations

While these findings provide valuable insights, it is important to acknowledge the study's limitations. One key limitation is that this study does not account for individual differences in how students engage with feedback over time. Future research could explore whether specific learner characteristics—such as motivation, or learning style—moderate the effectiveness of different feedback types. Additionally, investigating how students interact with automated feedback could provide deeper insights into how these tools contribute to long-term writing improvement.

The study examined the effect of a specific implementation of Criterion—an AWE tool—to provide students with automated feedback. This focus on a single AWE system may limit the generalizability of the findings, as different AWE tools vary in their feedback algorithms, error detection capabilities, and user interfaces. Future research could address this limitation by examining multiple AWE tools to determine whether feedback effectiveness varies across different platforms.

Another limitation is the sample size, as only 74 participants were involved in the study. However, given that three entire classes were recruited, the study maintains a degree of ecological validity, as it reflects a realistic classroom environment.

Moreover, only 32 participants completed the delayed posttest, which was largely due to post-pandemic restrictions that limited participation. Additionally, the delayed posttest was conducted five weeks after the intervention, coinciding with the start of a new semester. As students were assigned different schedules, tracking and ensuring their continued participation became more challenging. Future research may consider adopting different follow-up approaches to reduce participant dropout and improve data collection for long-term evaluations.

Finally, the specific population consisted of pre-intermediate foundation year students at the University of Jeddah. As a result, the findings may be difficult to generalize to other EFL programs with different demographics, pedagogical orientations, or institutional purposes.

Despite these limitations, there were several strengths. The study is one of the few to examine the hybrid approach to feedback and compare it to both AWE and teacher feedback simultaneously, rather than evaluating AWE solely as a replacement for teacher feedback. Also, the use of intact classes boosted the ecological validity of the study and provided a naturalistic setting for research.

7.4 Future Research Directions

This study has yielded insights into the effects of teacher-only, automated-only, and hybrid feedback on EFL learners' writing. To build directly on these findings, future research could investigate feedback uptake and retention over longer periods and with larger, more diverse samples. Such studies would provide more robust evidence on the sustainability of feedback effects and allow for stronger generalizability. Implementing randomized or quasi-experimental designs where feasible would further strengthen internal validity.

A key avenue for follow-up research is examining how learners engage with different feedback types. Employing a mixed-methods approach that combines quantitative outcome measures with qualitative data—such as interviews, focus groups, or think-aloud protocols—would shed light on learners' perceptions, processing strategies, and the cognitive effort involved in responding to feedback. Teacher interviews could also provide insight into effective strategies for combining teacher and automated feedback in classroom contexts.

Moreover, although this study relied solely on human raters for scoring, future research could use a dual scoring approach, combining human and automated scoring. Such triangulation would provide a stronger basis for evaluating writing performance and would allow researchers to explore whether human and computer scores diverge in systematic ways across writing genres or feedback conditions.

Drawing on the Cognition Hypothesis (Robinson, 2001), future research could also explore how feedback interacts with task complexity and genre-specific demands. Investigating whether certain feedback modes are more effective for cognitively demanding tasks—such as persuasive or argumentative writing—would deepen our understanding of the feedback-task relationship. Future research is encouraged to adopt robust validation methods to ensure that task complexity is empirically supported and accurately reflects cognitive demands.

Given the rapid advancement of feedback technologies, it would also be valuable to examine how emerging AWE tools, with increasingly sophisticated natural language processing capabilities, compared to current systems like Criterion. Future research might consider how newer tools influence feedback quality, learner engagement, and writing development. Moreover, future research could explore the effects of generative artificial intelligence (AI) feedback in comparison to traditional teacher feedback. As AI-based feedback tools continue to evolve, examining their impact on student writing and learning outcomes will provide valuable insights for both research and pedagogical practice.

Additionally, studies could explore how AWE systems are integrated into teacher training programs and curriculum development. Understanding how instructors interpret and utilize AWE data, and how their feedback practices evolve when working with these tools, would offer practical insights for sustainable classroom implementation.

In closing, this study has been both an academic and personal journey toward understanding how feedback, in its various forms, can shape the writing development of EFL learners. By exploring teacher, automated, and hybrid feedback within real classroom contexts, the research sheds light on practical ways to support learners more effectively and sustainably. It is my hope that these findings will not only contribute to ongoing scholarly dialogue but also inspire teachers, curriculum designers, and researchers to continue exploring innovative feedback approaches that empower learners. As writing instruction continues to evolve alongside technological advancement, the challenge—and opportunity—lies in creating responsive, inclusive, and effective feedback systems that meet the diverse needs of learners around the world.

Appendices

Appendix A-1: analytic rubric adapted from Connor-Linton, J., & Polio, C. (2014).

Score	Text content	Score	Text organization	Score	Vocabulary use	Score	Language use	Score	Mechanics
20 16	Thorough and logical development of thesis. Substantive and detailed. No irrelevant information. Interesting. A substantial number of words for the amount of time given.	20 16	Excellent overall organization. Clear thesis statement. Substantive introduction and conclusion. Excellent use of transition words. Excellent connections between paragraphs. Unity within every paragraph.	20 16	Very sophisticated vocabulary. Excellent choice of words with no errors. Excellent range of vocabulary. Idiomatic and near native-like vocabulary. Academic register.	20 16	No major errors in word order or complex structures. No errors that interfere with comprehension. Only occasional errors in morphology. Frequent use of complex sentences. Excellent sentence variety.	20 16	Appropriate layout with indented paragraphs No spelling errors. No punctuation errors.
15 11	Good and logical development of thesis. Fairly substantive and detailed. Almost no irrelevant information. Somewhat interesting. An adequate number of words for the time given.	15 11	Good overall organization. Clear thesis statement. Good introduction and conclusion. Good use of transition words. Good connections between paragraphs. Unity within most paragraphs.	15 11	Somewhat sophisticated vocabulary. Attempts, even if not completely successful, at sophisticated vocabulary. Good choice of words with some errors that don't obscure the meaning. Adequate range of vocabulary but some repetition. Approaching academic register.	15 11	Occasional errors in awkward order or complex structures. Almost no errors that interfere with comprehension. Attempts, even if not completely successful, at a variety of complex structures. Some errors in morphology. Frequent use of complex sentences. Good sentence variety.	15 11	Appropriate layout with indented paragraphs. No more than few spelling errors in less frequent vocabulary. No more than a few punctuation errors.
10	Some development of thesis.	10	Some general coherent organization.	10	Unsophisticated vocabulary .	10	Errors in word order or complex structures.	10	Appropriate layout with indented paragraphs.

6	Not much substance or details. Some irrelevant information. Somewhat uninteresting. Limited number of words for the amount of time given.	6	Minimal thesis statement or main idea. Minimal introduction and conclusion. Occasional use of transition words. Some disjointed connections between paragraphs. Some paragraphs may lack unity.	6	Limited word choice with some errors obscuring meaning. Repetitive choice of words. No resemblance to academic register.	6	Some errors that interfere with comprehension. Frequent errors in morphology. Minimal use of complex sentences. Little sentence variety.	6	Some spelling errors with less frequent vocabulary. Several punctuation errors.
5 0	No development of thesis. No substance or details. Substantial amount of irrelevant information. Completely uninteresting. Very few words for the amount of time given.	5 0	No coherent organization. No thesis statement or main idea. No introduction and conclusion. No use of transition words. Disjointed connections between paragraphs. Paragraph lacks unity.	5 0	Very simple vocabulary. Severe errors in word choice that often obscure the meaning. No variety in word choice. No resemblance to academic register.	5 0	Serious errors in word order or complex structures. Frequent errors that interfere with comprehension. Many errors in morphology. Almost no attempt at complex sentences. No sentence variety.	5 0	No attempt to arrange essay into paragraphs. Several spelling errors even in frequent vocabulary. Many punctuations errors.

Appendix A-2:

Student number: C529

Pretest	Content/20	Organization/20	Vocabulary/20	Language Use/20	Mechanics/20	Total/100
Writing Task1	10	10	10	10	10	50
Writing Task2	10	10	10	10	10	50

Posttest	Content/20	Organization/20	Vocabulary/20	Language Use/20	Mechanics/20	Total/100
Writing Task1	15	15	12	15	15	72
Writing Task2	10	15	10	15	5	55

Posttest

Student Code	C529
Writing task	1
Version	C
Group	

~~9.5~~ ~~10~~ A ~~16~~ C N

I'm going to be talking about the trolls, its an animated comedy film by dreamworks animation. The trolls are tiny, colorful, always happy creatures who sing and dance and hug all day the main character is called princes poppy, She is the daughter of King Peppy, later the queen of the trolls, poppy has a light pink skin with glitter freckles her hair color is darker pink, her eyebrows are the same color, and so are her nose and cheeks, her main outfit is a blue dress, and she wears a purple bracelet, she is the very positive leade of the trolls. She's kind, energetic, enthusiastic, and commited to helping everyone. She loves to hug, and she never gives up when someone is in need, the messages she sends are positive and easy to understand. (happiness is inside everyone, if you know where and how to find it, and you shouldnt have to change who you are to get someone to like you).

Student number: C529

Pretest	Content/20	Organization/20	Vocabulary/20	Language Use/20	Mechanics/20	Total/100
Writing Task1	10	10	10	10	10	50
Writing Task2	10	10	10	10	10	50

Posttest	Content/20	Organization/20	Vocabulary/20	Language Use/20	Mechanics/20	Total/100
Writing Task1	15	15	12	15	15	72
Writing Task2	10	15	10	15	5	55

Pretest

Student Code	C529
Writing task	1Schools
Version	B
Group	

encreate knoled
start a career

Studying is really important nowadays, to start a career, experience more stuff in life or even socialize more! Studying has improved so many benefits in s many ways, such as encreasing knowledge.

Most of people attend shcools by the age of five, they start learning letters and numbers and basics of other subjects, it is really important for them to have a foundaton and grow up to build on top of

Student number: D728

Pretest	Content/20	Organization/20	Vocabulary/20	Language Use/20	Mechanics/20	Total/100
Writing Task1	15	15	12	12	15	69
Writing Task2	15	15	15	15	15	75

Posttest	Content/20	Organization/20	Vocabulary/20	Language Use/20	Mechanics/20	Total/100
Writing Task1	18	18	15	15	15	81
Writing Task2	18	18	18	18	20	92

Post

Student Code	D728
Writing task	Task 2
Version	A
Group	

Teenage or Adult Life

Some people think that the teenage years are the happiest times of most people's lives. On the other side, other think that adult life brings more happiness. However, I will discuss how the two views.

For the first view, no one can deny that teenage life is full of challenges and fun. Teenagers can do whatever they want without being afraid of taking responsibility. All what they have to do is study... Also, they don't have to take care of a family or children.

On the contrary, adult life has more complicated things and responsibilities compared to the teenage years. Adults have to work well to earn money for their families. Moreover, they have to take care of their children and make sure that they don't need anything. Further, they must take responsibility for their teenagers' actions. They work everyday to provide a good life for their children.

In the end, each life of them has a good and bad side. In my opinion, teenage years are the best years in most people's lives.

Student number: C537

Pretest	Content/20	Organization/20	Vocabulary/20	Language Use/20	Mechanics/20	Total/100
Writing Task1	10	10	10	10	5	45
Writing Task2	8	8	5	8	5	34

Posttest	Content/20	Organization/20	Vocabulary/20	Language Use/20	Mechanics/20	Total/100
Writing Task1	13	11	11	15	10	60
Writing Task2	15	15	15	15	15	75

Posttest

Student Code	C537
Writing task	2
Version	B
Group	

we gain our knowledge about our lives and the world around us from two sources: from books and from experience. These two resources are both important; however, most of our knowlege is based on our participation in real life situations.
of course reading books from a wide range of our knowledge. we acquire many knowledge about law, politics, sociology, and physics when we read. we could use other peoples informations and experience that it is very difficult or impossible to do by with reading great stories of real peoples live, we can get many experience about how the solve thier daily success in the Futher. These are valuable lesson an individual needs to learn in this lifetime.
In conclusion, both learning sources, books and experience, are important for us. But in my opinion knowledge from experience is more important.
Experience is practical way to find theoretical knowledge that gained from books is real or not. In addition, many important aspect of our pessonality such as our feelings and skills develop through experince.

Appendix B:

B-1 List of categories and subcategories of *Criterion* feedback.

Feedback categories	Feedback subcategories	Example
Grammar	Fragments	*Boating around the lake when the storm moved in from the west.*
	Run-on Sentences	*As usual, the students are staging a school play this year, one of them also wrote the play.*
	Subject–verb Agreement	*The football players is holding a pizza.*
	Ill-formed Verbs	*He will learn to drove when he turns sixteen.*
	Pronoun Errors	*My friend and me signed up for the drama club.*
	Possessive Errors	*Toms' notebook was neat.*
	Wrong or Missing Word	*The decided to talk to the administration.*
Usage	Determiner Noun Agreement	*All player.*
	Wrong Article	*a elephant, an book,…etc.*
	Missing or Extra Article	
	Confused Word	*accept/except, advise/advice, affect/effect…etc.*
	Wrong Form of Word Faulty	*I am capability of studying before dinner.*
	Comparisons	*more braver, most tallest ,…etc.*
	Nonstandard Word Form	*I kinda like dancing.*
	Negation Error	*I am not going to pay no bills today.*
	Wrong Part of Speech	*The jet will flight from here.*
Mechanics	Spelling	*They will receve a certificate for attending.*
	Capitalize Proper Noun	*She will meet her friend noah at the shop.*
	Missing Initial Capital Letter	*they are running a marathon.*
	Missing Question Mark	*Where have you gone_*
	Missing Final Punctuation	*They are reading an interesting book_*

	Missing Apostrophe	*Theyre about to leave.*
	Missing Comma	*On Tuesday I take karate lessons.*
	Hyphen Error	*Missing hyphens in words like: hard-working, one-fifth , self-motivated … Anyother class is easier than chemistry.*
	Fused Words	*I told my self that I would finish reading this chapter today.*
	Compound Words	*I will can apply to the foreign study program in summer.*
	Duplicates	*I have a cat, and a dog.*
	Extra Comma	
Style	Repetition of Words	***Swimming*** is the best form of exercise because ***swimming*** gives you a good workout ….
	Inappropriate Words or Phrases	Using profanity, vulgar language or phrases that have the potential to offend readers.
	Sentences Beginning with Coordinating Conjunction	when too many sentences begin with coordinating conjunctions, your writing will appear fragmented and sound choppy.
	Short or long Sentences	Well-written essays feature sentences of varying lengths to make the writing more interesting and energetic. Using too many short sentences can make your ideas sound overly simplistic and your writing sound choppy.
	Passive or active voice	*Two juniors won the national debating tournament. / The national debating tournament was won by two juniors.*

Note: examples and explanations are taken from Criterion writer's handbook which is accessible to students who are using the software.

B – 2 Example of feedback provided in Criterion and used in teachers' training booklet to familiarize them with automatically generated content and organization feedback, which teachers will provide for students in teacher-only and hybrid feedback groups.

Content Feedback:

1- Development of the thesis.
2- Relevance of provided information.
3- Interesting or not.
4- Number of words.

Word Choice: vocabulary used.

Conventions: This refers to the grammar, mechanics and usage at the sentence level.

Fluency/Organization: This refers to the response as a whole. The level is based on the general structure (introduction, thesis, main points, supporting ideas, and conclusion), appropriate transitions, sentence variety, and proper use of active and passive voice.

Examples automatically generated by *Criterion*:



Appendix C: **Pre, Post, and Delayed tests prompts.**

VERSION A

WRITING TASK 1

You should spend ***20 minutes*** on this task.

Reasons for Attending College (Expository)

People attend a college or university for many different reasons (for example, new experiences, career preparation and increased knowledge). Why do you think people attend college or university? Use specific reasons and examples to support

Write at least 150 words.

WRITING TASK 2

You should spend ***40 minutes*** on this task.

Write about the following topic:

Teenage or Adult life (Persuasive)

Some people think that the teenage years are the happiest times of most people's lives. Others think that adult life brings more happiness, in spite of greater responsibilities.

Give reasons for your answer and include any relevant examples from your own knowledge or experience.

Write at least 250 words.

VERSION B

WRITING TASK 1

You should spend ***20 minutes*** on this task.

Why Study Abroad? (Expository)

Many students choose to attend schools or universities outside their home countries. Why do some students study abroad? Use specific reasons and details to explain your answer.

Write at least 150 words.

WRITING TASK 2

You should spend ***40 minutes*** on this task.

Write about the following topic:

Experience or Books (Persuasive)

It has been said, "Not everything that is learned is contained in books." Compare and contrast knowledge gained from experience with knowledge gained from books. In your opinion, which source is more important? Why?

Give reasons for your answer and include any relevant examples from your own knowledge or experience.

Write at least 250 words.

VERSION C

WRITING TASK 1

You should spend ***20 minutes*** on this task.

Fictional Character (Expository)

Fictional characters from any genre (whether in books, movies, video games, etc.) often prove to be unforgettable. Write an essay about any fictional character that has had an effect on you. Fully describe the character, where you discovered him or her, and the effect he or she has had on you.

Write at least 150 words.

WRITING TASK 2

You should spend ***40 minutes*** on this task.

Write about the following topic:

Learning A New Language (Persuasive)

People who are learning a foreign language can face a number of difficulties. What are some of these problems? In your opinion, what are the best ways to overcome these difficulties?

Give reasons for your answer and include any relevant examples from your own knowledge or experience.

Write at least 250 words.

Appendix D: Treatment sessions task prompts.

Session 1

Good Friend (Expository)

What are the qualities of a good friend? Write an essay in which you describe what it takes to be a good friend. Identify the qualities a person must have to be a good friend, and develop those ideas with specific examples and support, citing your own experiences.

Session 2

Money on Technology (Persuasive)

Some people think that governments should spend as much money as possible on developing or buying computer technology. Other people disagree and think that this money should be spent on more basic needs. Which one of these opinions do you agree with? Use specific reasons and details to support your answer.

Session 3

New Product (Expository)

If you could invent something new, what product would you develop? Use specific details to explain why this invention is needed.

Session 4

Change Job or Not (Persuasive)

Some people prefer to change jobs or professions during their careers. Others choose to stay in the same job or profession. Discuss the advantages of each choice. Which do you prefer? Use reasons and examples to explain your choice.

Appendix E: **QQ plots for overall writing production.**

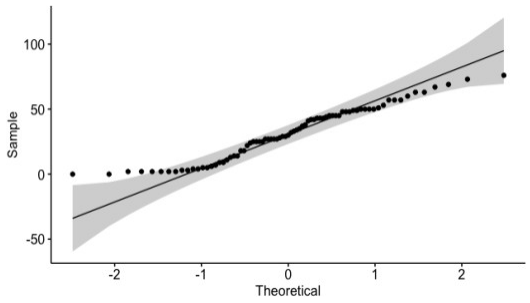


Figure E.1 *Pretest scores Expository Genre*

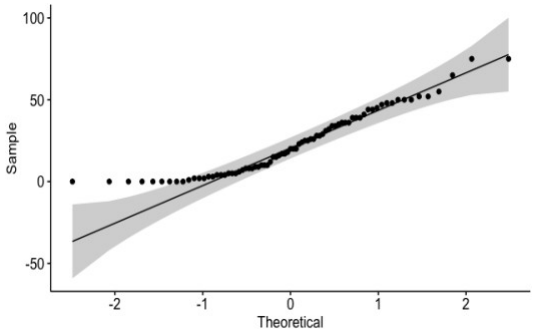


Figure E.2 *Pretest scores Persuasive Genre*

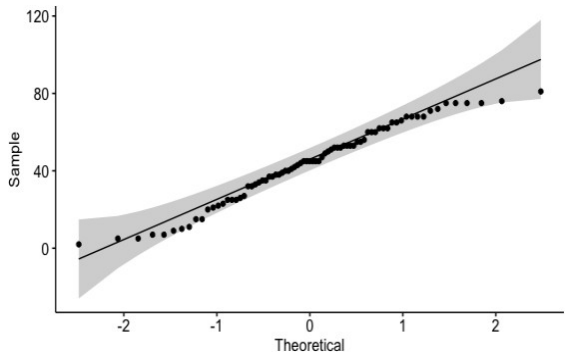


Figure E.3 *Posttest scores Expository Genre*

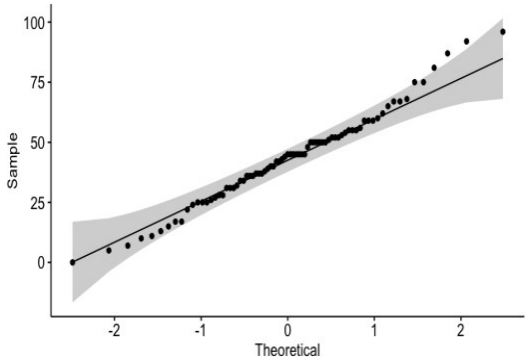


Figure E.4 *Posttest scores Persuasive Genre*

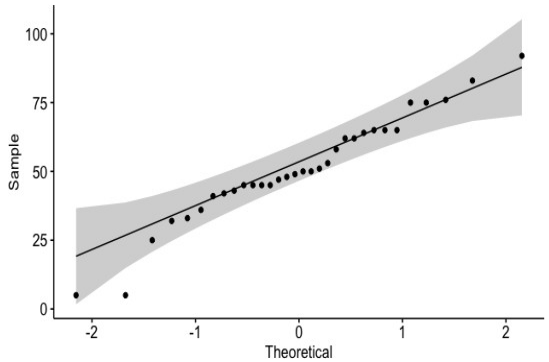


Figure E.5 *Delayed Posttest Expository Genre.*

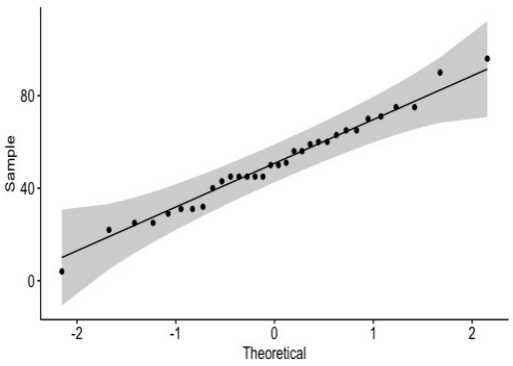


Figure E.6 *Delayed Posttest Persuasive Genre*

Appendix F: **Histograms for overall writing production.**

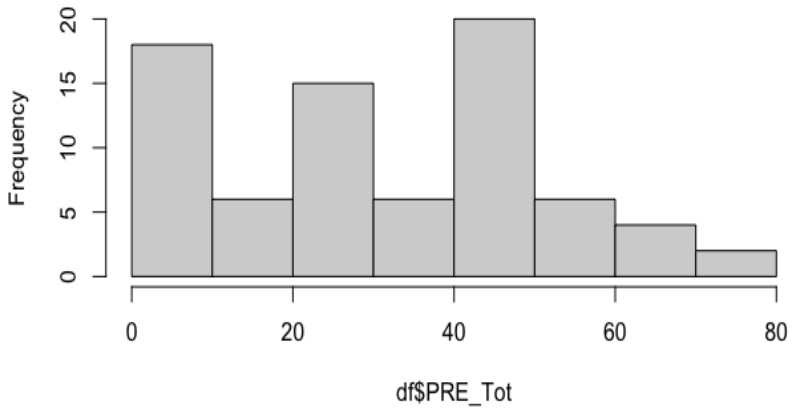


Figure F.1 *Pretest Expository Genre*

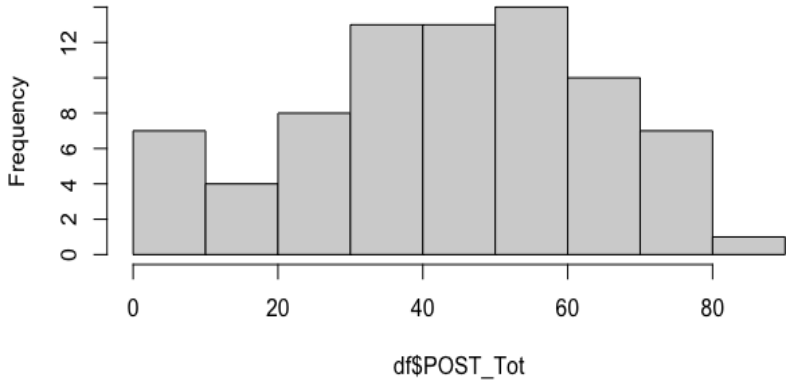


Figure F.2 *Posttest Expository Genre*

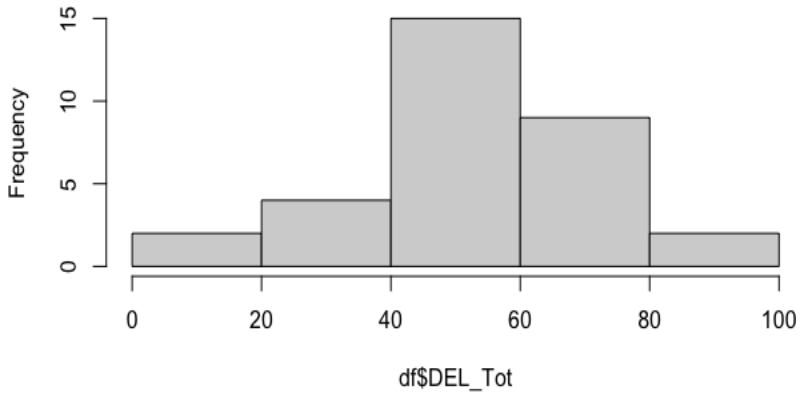


Figure F.3 *Delayed Posttest Expository Genre*

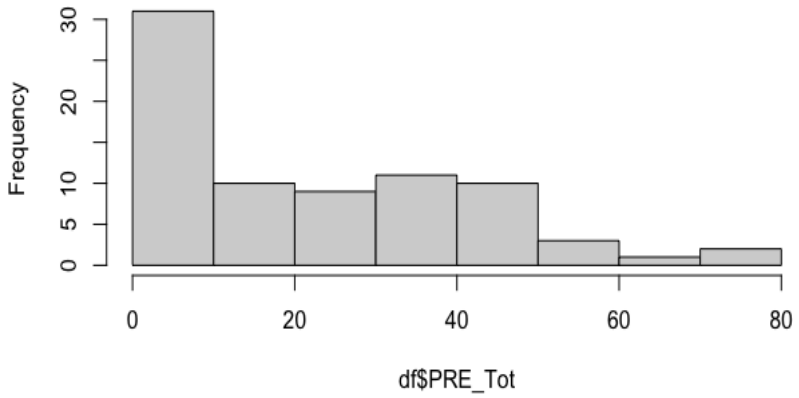


Figure F.4 *Pretest Persuasive Genre*

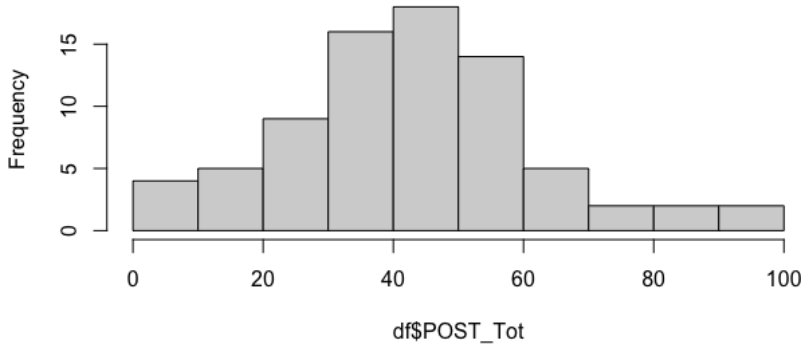


Figure F.5 *Posttest Expository Genre*

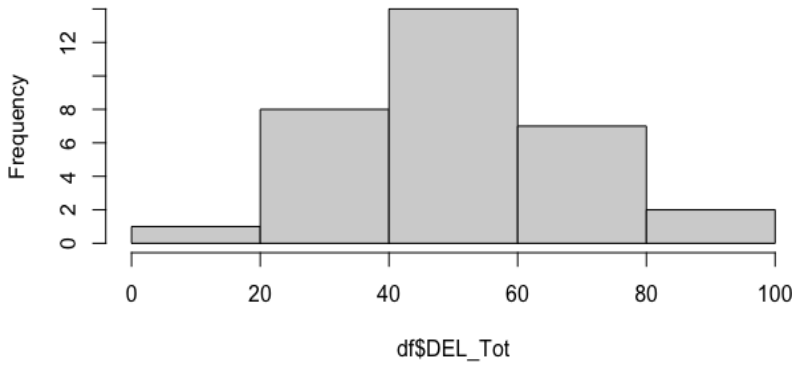


Figure F.6 *Delayed Posttest Persuasive Genre*

Appendix G: **Mixed-effects Model Outcome**

G.1 *Summary of Fixed Effect Predictors of Mixed-effects Model Outcome for Total Writing Production Data with Teacher Feedback Group as a Reference.*

Reference level	Predictors	Estimates [95% CIs]	*SE*	*df*	*t*	*P*
Teacher feedback group & pretest	Intercept	49.13 [33.71 – 64.54]	7.85	5.59	6.26	0.0010**
	Posttest	37.72 [32.95 – 42.49]	2.43	820.12	15.54	<2.00E-16***
	Delayed posttest	36.85 [30.45 – 43.24]	3.26	829.83	11.31	<2.00E-16***
	AF group	14.30 [-2.34 – 30.95]	8.48	77.70	1.69	0.0957
	ATF group	8.73 [-7.94 – 25.39]	8.49	76.99	1.03	0.30745
	Posttest: AF	-4.16 [-10.90 – 2.58]	3.43	820.12	-1.21	0.22605
	Delayed: AF	11.45 [2.02 – 20.87]	4.80	831.08	2.38	0.01735
	Posttest: ATF	-10.80 [-17.61 – - 3.99]	3.47	820.12	-3.11	0.00191*
	Delayed: ATF	21.89 [12.46 – 31.33]	4.81	831.26	4.55	6.06E-06***
Teacher feedback group & posttest	Intercept	86.85[71.43 – 102.26]	7.85	5.59	11.06	5.19E-05***
	Pretest	-37.72[-42.49 – -32.95]	2.43	820.12	-15.54	2.00E-16***
	Delayed posttest	-0.87[-7.27 – 5.52]	3.26	829.83	-0.27	0.78845
	AF group	10.14[-6.50 – 26.79]	8.48	77.70	1.20	0.23533
	ATF group	-2.08[-18.75 – 14.59]	8.49	76.99	-0.25	0.8073
	Pretest: AF	4.16[-2.58 – 10.90]	3.43	820.12	1.21	0.22605
	Delayed: AF	15.61[6.18 – 25.03]	4.80	831.08	3.25	0.0012**
	Pretest: ATF	10.80[3.99 – 17.61]	3.47	820.12	3.11	0.00191**
	Delayed: ATF	32.69[23.26 – 42.13]	4.81	831.26	6.80	1.98E-11***
Teacher feedback group & delayed test	Intercept	85.97[69.98 – 101.97]	7.85	5.59	11.06	2.55E-05***
	Pretest	-36.85[-43.24 – -30.45]	2.43	820.12	-15.54	2.00E-16***
	Posttest	0.87[-5.52 – 7.27]	3.26	829.83	-0.27	0.78845
	AF group	25.75[7.84 – 43.66]	8.48	77.70	1.20	0.00573*
	ATF group	30.62[12.72 – 48.52]	8.49	76.99	-0.25	0.00111**
	Pretest: AF	-11.45[-20.87 – -2.02]	3.43	820.12	1.21	0.01735
	Posttest: AF	-15.61[-25.03 – -6.18]	4.80	831.08	3.25	0.0012**
	Pretest: ATF	-21.89[-31.33 – -12.46]	3.47	820.12	3.11	6.06E-06***
	Posttest: ATF	-32.69[-42.13 – -23.26]	4.81	831.26	6.80	1.98E-11***

Note. Alpha values adjusted to correct for four repeated tests. *SE* = standard error; *df* = degree of freedom; *AF* = automated feedback group; *ATF* = automated+teacher feedback group.

*$p < .0125$. **$p < .0025$. ***$p < .00025$

G.2 *Summary of fixed effect predictors of mixed-effects model outcome for total writing production data with Automated feedback group as a reference.*

Reference level	Predictors	Estimates [95% CIs]	*SE*	*df*	*t*	*P*
Automated feedback group & pretest	Intercept	63.43[48.17 – 78.69]	7.78	5.55	8.16	0.00027**
	Posttest	33.56[28.79 – 38.33]	2.43	820.12	13.82	2.00E-16***
	Delayed posttest	48.29[41.37 – 55.21]	3.53	832.35	13.70	2.00E-16***
	TF group	-14.30[-30.95 – 2.34]	8.48	77.70	-1.69	0.096
	ATF group	-5.58[-22.20 – 11.04]	8.47	76.66	-0.66	0.512
	Posttest:TF	4.16[-2.58 – 10.90]	3.43	820.12	1.21	0.226
	Delayed:TF	-11.45[-20.87 – -2.02]	4.80	831.08	-2.38	0.017
	Posttest: ATF	-6.64[-13.45 – 0.17]	3.47	820.12	-1.92	0.0558
	Delayed: ATF	10.45[0.65 – 20.24]	4.99	832.03	2.09	0.036
Automated feedback group & posttest	Intercept	96.99[81.73 – 112.25]	7.78	5.6	12.48	2.9E-05***
	Pretest	-33.56[-38.33 – -28.79]	2.43	820.1	-13.82	2.0E-16***
	Delayed posttest	14.73[7.81 – 21.65]	3.53	832.4	4.18	3.25E-05***
	TF group	-10.14[-26.79 – 6.50]	8.48	77.7	-1.20	0.23533
	ATF group	-12.22[-28.84 – 4.40]	8.47	76.7	-1.44	0.15308
	Pretest: TF	-4.16[-10.90 – 2.58]	3.43	820.1	-1.21	0.22605
	Delayed: TF	-15.61[-25.03 – -6.18]	4.80	831.1	-3.25	0.0012**
	Pretest: ATF	6.64[-0.17 – 13.45]	3.47	820.1	1.92	0.05585
	Delayed: ATF	17.09[7.29 – 26.89]	4.99	832.0	3.42	0.00065**
Automated feedback group & delayed test	Intercept	111.72[95.66 – 127.79]	8.19	6.81	13.65	3.4E-06***
	Pretest	-48.29[-55.21 – -41.37]	3.53	832.35	-13.70	2.0E-16***
	Posttest	-14.73[-21.65 – -7.81]	3.53	832.35	-4.18	3.3E-05***
	TF group	-25.75[-43.66 – -7.84]	9.12	103.31	-2.82	0.00573*
	ATF group	4.87[-13.19 – 22.93]	9.20	105.83	0.53	0.598
	Pretest: TF	11.45[2.02 – 20.87]	4.80	831.08	2.38	0.01735
	Posttest:TF	15.61[6.18 – 25.03]	4.80	831.08	3.25	0.0012**
	Pretest: ATF	-10.45[-20.24 – -0.65]	4.99	832.03	-2.09	0.03673
	Posttest: ATF	-17.09[-26.89 – -7.29]	4.99	832.03	-3.42	0.00065***

Note. Alpha values adjusted to correct for four repeated tests. *SE* = standard error; *df* = degree of freedom; *TF* = teacher feedback group; *ATF* = automated+teacher feedback group.

*$p < .0125$. **$p < .0025$. ***$p < .00025$

G.3 *Summary of fixed effect predictors of mixed-effects model outcome for total writing production data with Automated+teacher feedback as a reference.*

Reference level	Predictors	Estimates [95% CIs]	*SE*	*df*	*t*	*P*
Automated + teacher feedback group & pretest	Intercept	57.85 [42.41 – 73.29]	7.87	5.81	7.35	0.00038**
	Posttest	26.92 [22.05 – 31.78]	2.48	820.12	10.86	2.0E-16***
	Delayed posttest	58.74 [51.80 – 65.67]	3.54	831.85	16.62	2.0E-16***
	TF group	-8.73 [-25.39 – 7.94]	8.49	76.99	-1.03	0.307
	AF group	5.58 [-11.04 – 22.20]	8.47	76.66	0.66	0.512
	Posttest: TF	10.80 [3.99 – 17.61]	3.47	820.12	3.11	0.002**
	Delayed: TF	-21.89 [-31.33 – -12.46]	4.81	831.26	-4.55	6.1E-06***
	Posttest: AF	6.64 [-0.17 – 13.45]	3.47	820.12	1.92	0.055
	Delayed: AF	-10.45 [-20.24 – -0.65]	4.99	832.03	-2.09	0.037
Automated + teacher feedback group & posttest	Intercept	84.77 [69.33 – 100.21]	7.87	5.81	10.78	4.6E-05***
	Pretest	-26.92 [-31.78 – -22.05]	2.48	820.12	-10.86	2.0E-16***
	Delayed posttest	31.82 [24.88 – 38.76]	3.54	831.85	9.00	2.0E-16***
	TF group	2.08 [-14.59 – 18.75]	8.49	76.99	0.25	0.8073
	AF group	12.22 [-4.40 – 28.84]	8.47	76.66	1.44	0.15308
	Pretest: TF	-10.80 [-17.61 – -3.99]	3.47	820.12	-3.11	0.002**
	Delayed: TF	-32.7 [-42.13 – -23.26]	4.81	831.26	-6.80	1.98E-11***
	Pretest: AF	-6.64 [-13.45 – 0.17]	3.47	820.12	-1.92	0.05585
	Delayed: AF	-17.09 [-26.89 – -7.29]	4.99	832.03	-3.42	0.00065**
Automated + teacher feedback group & delayed test	Intercept	116.8 [100.53 – 133.03]	8.28	7.13	14.11	1.8E-06***
	Pretest	-59.24 [-66.24 – -52.23]	3.57	575.63	-16.60	2.0E-16***
	Posttest	-31.75 [-38.69 – -24.81]	3.54	831.80	-8.98	2.0E-16***
	TF group	-30.81 [-48.72 – -12.90]	9.12	101.94	-3.38	0.0010**
	AF group	-5.19 [-23.28 – 12.89]	9.22	106.54	-0.56	0.574
	Pretest: TF	22.32 [12.82 – 31.83]	4.84	654.00	4.61	4.9E-06***
	Posttest: TF	32.73 [23.25 – 42.21]	4.83	758.80	6.78	2.44E-11***
	Pretest: AF	11.09 [1.17 – 21.01]	5.06	469.71	2.19	0.0287
	Posttest: AF	17.14 [7.24 – 27.04]	5.04	556.25	3.40	0.00072**

Note. Alpha values adjusted to correct for four repeated tests. *SE* = standard error; *df* = degree of freedom.

TF = teacher feedback group; *AF* = automated feedback group.

*$p < .0125$. **$p < .0025$. ***$p < .00025$

G.4 *Summary of fixed effect predictors of mixed-effects model outcome for Text content score with Teacher feedback group as a reference.*

Reference level	Predictors	Estimates [95% CIs]	*SE*	*df*	*t*	*P*
Teacher feedback group & pretest	Intercept	11.16 [7.78 – 14.54]	1.71	9.94	6.52	6.94E-05***
	Posttest	7.36 [4.45 – 10.27]	1.47	101.45	5.00	2.45E-06***
	Delayed posttest	6.26 [2.45 – 10.08]	1.93	111.48	3.24	0.00158**
	AF group	2.97 [-1.11 – 7.05]	2.07	126.82	1.44	0.15361
	ATF group	1.92 [-2.18 – 6.02]	2.08	126.26	0.92	0.35725
	Posttest: AF	-0.8 [-4.91 – 3.31]	2.08	101.45	-0.38	0.70178
	Delayed: AF	1.38 [-4.23 – 6.99]	2.84	112.86	0.48	0.62908
	Posttest: ATF	-2.11 [-6.27 – 2.05]	2.11	101.45	-1.00	0.31853
	Delayed: ATF	5.99 [0.38 – 11.61]	2.84	112.99	2.11	0.03739
Teacher feedback group & posttest	Intercept	18.52 [15.14 – 21.90]	1.71	9.94	10.81	8.14E-07***
	Pretest	-7.36 [-10.27 – -4.45]	1.47	101.45	-5.00	2.45E-06***
	Delayed posttest	-1.09 [-4.91 – 2.72]	1.93	111.48	-0.57	0.57256
	AF group	2.17 [-1.91 – 6.25]	2.07	126.82	1.05	0.29634
	ATF group	-0.19 [-4.29 – 3.91]	2.08	126.26	-0.09	0.92704
	Pretest: AF	0.8 [-3.31 – 4.91]	2.08	101.45	0.38	0.70178
	Delayed: AF	2.18 [-3.43 – 7.79]	2.84	112.86	0.77	0.44536
	Pretest: ATF	2.11 [-2.05 – 6.27]	2.10	101.45	1.00	0.31853
	Delayed: ATF	8.10 [2.49 – 13.72]	2.84	112.99	2.85	0.00523*
Teacher feedback group & delayed test	Intercept	17.43 [13.24 – 21.62]	2.12	22.14	8.21	3.65E-08***
	Pretest	-6.27 [-10.08 – -2.45]	1.93	111.48	-3.24	0.00158**
	Posttest	1.09 [-2.72 – 4.91]	1.93	111.48	0.57	0.57256
	AF group	4.34 [-1.25 – 9.94]	2.84	170.70	1.53	0.12721
	ATF group	7.91 [2.34 – 13.48]	2.82	169.17	2.80	0.00567*
	Pretest: AF	-1.38 [-6.99 – 4.23]	2.84	112.86	-0.48	0.62908
	Posttest: AF	-2.18 [-7.79 – 3.43]	2.84	112.86	-0.77	0.44536
	Pretest: ATF	-5.99 [-11.61 – -0.38]	2.84	112.99	-2.11	0.03739
	Posttest: ATF	-8.10 [-13.72 – -2.49]	2.84	112.99	-2.85	0.00523*

Note. Alpha values adjusted to correct for four repeated tests. *SE* = standard error; *df* = degree of freedom; *AF* = automated feedback group; *ATF* = automated+teacher feedback group.

*$p < .0125$. **$p < .0025$. ***$p < .00025$

G.5 *Summary of fixed effect predictors of mixed-effects model outcome Text content score with Automated feedback group as a reference.*

Reference level	Predictors	Estimates [95% CIs]	*SE*	*df*	*t*	*P*
Automated feedback group & pretest	Intercept	14.13 [10.78 – 17.48]	1.70	10.11	8.32	7.81E-06***
	Posttest	6.56 [3.65 – 9.47]	1.47	101.45	4.45	2.18E-05***
	Delayed posttest	7.64 [3.53 – 11.75]	2.08	114.26	3.67	0.000367**
	TF group	-2.97 [-7.05 – 1.11]	2.07	126.82	-1.44	0.153607
	ATF group	-1.05 [-5.14 – 3.04]	2.07	125.86	-0.51	0.613821
	Posttest:TF	0.8 [-3.31 – 4.91]	2.08	101.45	0.38	0.701785
	Delayed:TF	-1.38 [-6.99 – 4.23]	2.84	112.86	-0.48	0.62908
	Posttest: ATF	-1.31 [-5.47 - 2.85]	2.11	101.45	-0.62	0.535112
	Delayed: ATF	4.62 [-1.21 – 10.44]	2.95	113.90	1.57	0.120341
Automated feedback group & posttest	Intercept	20.69 [17.34 – 24.04]	1.70	10.11	12.18	2.29E-07***
	Pretest	-6.56 [-9.47 – -3.65]	1.47	101.45	-4.45	2.18E-05***
	Delayed posttest	1.08 [-3.03 – 5.1]	2.08	114.26	0.52	0.6042
	TF group	-2.17 [-6.25 – 1.91]	2.07	126.82	-1.05	0.2963
	ATF group	-2.36 [-6.45 – 1.73]	2.07	125.86	-1.14	0.2574
	Pretest: TF	-0.8 [-4.91 – 3.31]	2.08	101.45	-0.38	0.7018
	Delayed: TF	-2.18 [-7.79 – 3.43]	2.84	112.86	-0.77	0.4454
	Pretest: ATF	1.31 [-2.85 – 5.47]	2.11	101.45	0.62	0.5351
	Delayed: ATF	5.93 [0.10 – 11.75]	2.95	113.90	2.01	0.0469
Automated feedback group & delayed test	Intercept	21.77 [17.34 – 26.21]	2.25	28.44	9.69	1.64E-10***
	Pretest	-7.64 [-11.75 – -3.53]	2.08	114.26	-3.67	.00037**
	Posttest	-1.08 [-5.19 – 3.03]	2.08	114.26	-0.52	0.604174
	TF group	-4.35 [-9.94 – 1.25]	2.84	170.70	-1.53	0.127205
	ATF group	3.57 [-2.22 – 9.35]	2.93	169.89	1.22	0.22516
	Pretest: TF	1.38 [-4.23 – 6.99]	2.84	112.86	0.48	0.62908
	Posttest:TF	2.18 [-3.43 – 7.79]	2.84	112.86	0.77	0.445355
	Pretest: ATF	-4.62 [-10.44 – 1.21]	2.95	113.90	-1.57	0.120341
	Posttest: ATF	-5.93 [-11.75 – -0.10]	2.95	113.90	-2.01	0.046865

Note. Alpha values adjusted to correct for four repeated tests. *SE* = standard error; *df* = degree of freedom; *TF* = teacher feedback group; *ATF* = automated+teacher feedback group.

*$p < .0125$. **$p < .0025$. ***$p < .00025$

G.6 *Summary of fixed effect predictors of mixed-effects model outcome for Text content score with Automated+teacher feedback as a reference.*

Reference level	Predictors	Estimates [95% CIs]	*SE*	*df*	*t*	*P*
Automated + teacher feedback group & pretest	Intercept	13.08 [9.68 – 16.49]	1.72	10.70	7.59	1.27E-05***
	Posttest	5.25 [2.28 – 8.22]	1.50	101.45	3.49	7.12E-04***
	Delayed posttest	12.26 [8.13 – 16.38]	2.09	113.69	5.87	4.43E-08***
	TF group	-1.92 [-6.02 – 2.18]	2.08	126.26	-0.92	0.357251
	AF group	1.05 [-3.04 – 5.14]	2.07	125.86	0.51	0.613821
	Posttest: TF	2.11 [-2.05 – 6.27]	2.11	101.45	1.00	0.318534
	Delayed: TF	-5.99 [-11.61 – -0.38]	2.84	112.99	-2.11	0.037393
	Posttest: AF	1.31 [-2.85 – 5.47]	2.11	101.45	0.62	0.535112
	Delayed: AF	-4.62 [-10.44 – 1.21]	2.95	113.90	-1.57	0.120341
Automated + teacher feedback group & posttest	Intercept	18.33 [14.93 – 21.74]	1.72	10.70	10.63	5.08E-07***
	Pretest	-5.25 [-8.22 – -2.28]	1.50	101.45	-3.49	7.12E-04***
	Delayed posttest	7.01 [2.88 – 11.13]	2.09	113.69	3.36	0.0012**
	TF group	0.19 [-3.91 – 4.29]	2.08	126.26	0.09	0.927
	AF group	2.36 [-1.73 – 6.45]	2.07	125.86	1.14	0.257
	Pretest: TF	-2.11 [-6.27 – 2.05]	2.10	101.45	-1.00	0.318
	Delayed: TF	-8.10 [-13.72 – -2.49]	2.84	112.99	-2.85	0.005**
	Pretest: AF	-1.31 [-5.47 – 2.85]	2.10	101.45	-0.62	0.535
	Delayed: AF	-5.92 [-11.75 – -0.10]	2.95	113.90	-2.01	0.047
Automated + teacher feedback group & delayed test	Intercept	25.34 [20.89 – 29.79]	2.25	28.71	11.25	4.92E-12***
	Pretest	-12.26 [-16.38 – -8.13]	2.09	113.69	-5.87	4.43E-08***
	Posttest	-7.01 [-11.13 – -2.88]	2.09	113.69	-3.36	0.0012**
	TF group	-7.91 [-13.48 – -2.34]	2.82	169.17	-2.80	0.0057*
	AF group	-3.57 [-9.35 – 2.22]	2.93	169.89	-1.22	0.225
	Pretest: TF	5.99 [0.38 – 11.61]	2.84	112.99	2.11	0.037
	Posttest: TF	8.10 [2.49 – 13.72]	2.84	112.99	2.85	0.005*
	Pretest: AF	4.62 [-1.21 – 10.44]	2.95	113.90	1.57	0.120
	Posttest: AF	5.93 [0.10 – 11.75]	2.95	113.90	2.01	0.047

Note. Alpha values adjusted to correct for four repeated tests. *SE* = standard error; *df* = degree of freedom.

TF = teacher feedback group; *AF* = automated feedback group.

*$p < .0125$. **$p < .0025$. ***$p < .00025$

G.7 *Summary of fixed effect predictors of mixed-effects model outcome for Text organization score with Teacher feedback group as a reference.*

Reference level	Predictors	Estimates [95% CIs]	*SE*	*df*	*t*	*P*
Teacher feedback group & pretest	Intercept	10.29 [6.49 – 14.09]	1.925	7.59	5.347	0.00082**
	Posttest	6.88 [3.70 – 10.06]	1.61	101.86	4.28	4.32E-05***
	Delayed posttest	7.45 [3.28 – 11.60]	2.11	113.02	3.53	0.0006**
	AF group	1.79 [-2.48 – 6.07]	2.17	132.28	0.83	0.409
	ATF group	1.96 [-2.33 – 6.26]	2.18	132.04	0.90	0.369
	Posttest: AF	1.64 [-2.85 – 6.13]	2.28	101.86	0.72	0.473
	Delayed: AF	2.39 [-3.72 – 8.50]	3.09	114.62	0.77	0.442
	Posttest: ATF	-1.38 [-5.92 – 3.16]	2.30	101.86	-0.60	0.549
	Delayed: ATF	4.8 [-1.31 – 10.91]	3.10	114.82	1.55	0.124
Teacher feedback group & posttest	Intercept	17.17 [13.37 – 20.97]	1.92	7.59	8.92	2.75E-05***
	Pretest	-6.88 [-10.06 – -3.70]	1.61	101.85	-4.28	4.32E-05***
	Delayed posttest	0.56 [-3.60 – 4.72]	2.11	113.02	0.27	0.789
	AF group	3.43 [-0.84 – 7.71]	2.17	132.28	1.59	0.116
	ATF group	0.58 [-3.71 – 4.88]	2.18	132.03	0.27	0.789
	Pretest: AF	-1.64 [-6.13 – 2.85]	2.28	101.85	-0.72	0.473
	Delayed: AF	0.75 [-5.36 – 6.86]	3.09	114.62	0.24	0.809
	Pretest: ATF	1.38 [-3.16 – 5.92]	2.30	101.85	0.60	0.549
	Delayed: ATF	6.18 [0.07 – 12.29]	3.10	114.82	2.00	0.048
Teacher feedback group & delayed test	Intercept	17.74 [13.08 – 22.39]	2.36	16.44	7.53	1.03E-06***
	Pretest	-7.44 [-11.60 – -3.28]	2.11	113.02	-3.53	.0006**
	Posttest	-0.56 [-4.72 – 3.60]	2.11	113.02	-0.27	0.789
	AF group	4.18 [-1.78 – 10.15]	3.02	170.66	1.39	0.168
	ATF group	6.76 [0.83 – 12.69]	3.00	169.38	2.25	0.026
	Pretest: AF	-2.39 [-8.50 – 3.72]	3.09	114.62	-0.77	0.442
	Posttest: AF	-0.75 [-6.86 – 5.36]	3.09	114.62	-0.24	0.809
	Pretest: ATF	-4.8 [-10.91 – 1.31]	3.10	114.82	-1.55	0.124
	Posttest: ATF	-6.18 [-12.29 – -0.07]	3.10	114.82	-2.00	0.048

Note. Alpha values adjusted to correct for four repeated tests. *SE* = standard error; *df* = degree of freedom; *AF* = automated feedback group; *ATF* = automated+teacher feedback group.

*$p < .0125$. **$p < .0025$. ***$p < .00025$

G.8 *Summary of fixed effect predictors of mixed-effects model outcome Text organization score with Automated feedback group as a reference.*

Reference level	Predictors	Estimates [95% CIs]	*SE*	*df*	*t*	*P*
Automated feedback group & pretest	Intercept	12.08 [8.32 – 15.85]	1.91	7.55	6.33	2.87E-04***
	Posttest	8.52 [5.34 – 11.7]	1.61	101.85	5.29	6.92E-07***
	Delayed posttest	9.83 [5.36 – 14.30]	2.26	116.27	4.34	3.02E-05***
	TF group	-1.79 [-6.07 – 2.48]	2.17	132.28	-0.83	0.409
	ATF group	0.17 [-4.12 – 4.45]	2.17	131.82	0.08	0.939
	Posttest:TF	-1.64 [-6.13 – 2.85]	2.28	101.85	-0.72	0.473
	Delayed:TF	-2.4 [-8.50 – 3.72]	3.09	114.62	-0.77	0.442
	Posttest: ATF	-3.02 [-7.56 – 1.52]	2.30	101.85	-1.31	0.192
	Delayed: ATF	2.41 [-3.92 – 8.75]	3.21	115.85	0.75	0.454
Automated feedback group & posttest	Intercept	20.60 [16.84 – 24.37]	1.91	7.55	10.79	7.40E-06***
	Pretest	-8.52 [-11.70 – -5.34]	1.61	101.85	-5.29	6.92E-07***
	Delayed posttest	1.31 [-3.16 – 5.78]	2.26	116.27	0.58	0.563
	TF group	-3.43 [-7.71 – 0.84]	2.17	132.28	-1.59	0.116
	ATF group	-2.85 [-7.14 – 1.43]	2.17	131.82	-1.31	0.191
	Pretest: TF	1.64 [-2.85 – 6.13]	2.28	101.85	0.72	0.473
	Delayed: TF	-0.75 [-6.86 – 5.36]	3.09	114.62	-0.24	0.809
	Pretest: ATF	3.02 [-1.52 – 7.56]	2.30	101.85	1.31	0.192
	Delayed: ATF	5.43 [-0.90 – 11.77]	3.21	115.85	1.69	0.093
Automated feedback group & delayed test	Intercept	21.92 [17.01 – 26.83]	2.49	20.62	8.81	1.97E-08***
	Pretest	-9.83 [-14.30 – -5.36]	2.26	116.27	-4.34	3.02E-05***
	Posttest	-1.31 [-5.78 – 3.16]	2.26	116.27	-0.58	0.563
	TF group	-4.18 [-10.15 – 1.78]	3.02	170.66	-1.39	0.168
	ATF group	2.58 [-3.58 – 8.74]	3.12	169.60	0.83	0.41
	Pretest: TF	2.39 [-3.72 – 8.50]	3.09	114.62	0.77	0.44
	Posttest:TF	0.75 [-5.36 – 6.86]	3.09	114.62	0.24	0.809
	Pretest: ATF	-2.41 [-8.75 – 3.92]	3.21	115.85	-0.75	0.454
	Posttest: ATF	-5.43 [-11.77 – 0.90]	3.21	115.85	-1.69	0.093

Note. Alpha values adjusted to correct for four repeated tests. *SE* = standard error; *df* = degree of freedom; *TF* = teacher feedback group; *ATF* = automated+teacher feedback group.

*$p < .0125$. **$p < .0025$. ***$p < .00025$

G.9 *Summary of fixed effect predictors of mixed-effects model outcome for Text organization score with Automated+teacher feedback as a reference.*

Reference level	Predictors	Estimates [95% CIs]	*SE*	*df*	*t*	*P*
Automated + teacher feedback group & pretest	Intercept	12.25 [8.43 – 16.07]	1.93	7.94	6.34	2.31E-04***
	Posttest	5.50 [2.26 – 8.74]	1.64	101.85	3.35	.0011**
	Delayed posttest	12.24 [7.76 – 16.73]	2.27	115.61	5.39	3.83E-07***
	TF group	-1.96 [-6.26 – 2.33]	2.18	132.03	-0.90	0.369
	AF group	-0.17 [-4.45 – 4.12]	2.17	131.82	-0.08	0.939
	Posttest: TF	1.38 [-3.16 – 5.92]	2.30	101.85	0.60	0.549
	Delayed: TF	-4.8 [-10.91 – 1.31]	3.10	114.82	-1.55	0.123
	Posttest: AF	3.02 [-1.52 – 7.56]	2.30	101.85	1.31	0.192
	Delayed: AF	-2.41 [-8.75 – 3.92]	3.21	115.85	-0.75	0.454
Automated + teacher feedback group & posttest	Intercept	17.75 [13.93 – 21.57]	1.93	7.94	9.18	1.68E-05***
	Pretest	-5.5 [-8.74 – -2.26]	1.64	101.86	-3.35	.0011**
	Delayed posttest	6.74 [2.26 – 11.23]	2.27	115.61	2.97	0.004*
	TF group	-0.58 [-4.88 – 3.71]	2.18	132.04	-0.27	0.790
	AF group	2.85 [-1.43 – 7.14]	2.17	131.82	1.31	0.191
	Pretest: TF	-1.38 [-5.92 – 3.16]	2.30	101.86	-0.60	0.550
	Delayed: TF	-6.18 [-12.29 – -0.07]	3.10	114.82	-2.00	0.048
	Pretest: AF	-3.02 [-7.56 – 1.52]	2.30	101.86	-1.31	0.192
	Delayed: AF	-5.43 [-11.77 – 0.90]	3.21	115.85	-1.69	0.093
Automated + teacher feedback group & delayed test	Intercept	24.50 [19.57 – 29.42]	2.49	20.78	9.83	2.92E-09***
	Pretest	-12.24 [-16.73 – -7.76]	2.27	115.61	-5.39	3.83E-07***
	Posttest	-6.74 [-11.23 – -2.26]	2.27	115.61	-2.97	0.004*
	TF group	-6.76 [-12.69 – -0.83]	3.00	169.38	-2.25	0.026
	AF group	-2.58 [-8.74 – 3.58]	3.12	169.60	-0.83	0.410
	Pretest: TF	4.8 [-1.31 – 10.91]	3.10	114.82	1.55	0.124
	Posttest: TF	6.18 [0.07 – 12.29]	3.10	114.82	2.00	0.048
	Pretest: AF	2.41 [-3.92 – 8.75]	3.21	115.85	0.75	0.454
	Posttest: AF	5.43 [-0.90 – 11.77]	3.21	115.85	1.69	0.093

Note. Alpha values adjusted to correct for four repeated tests. *SE* = standard error; *df* = degree of freedom.

TF = teacher feedback group; *AF* = automated feedback group.

*$p < .0125$. **$p < .0025$. ***$p < .00025$

G.10 *Summary of fixed effect predictors of mixed-effects model outcome for Vocabulary Use score with Teacher feedback group as a reference.*

Reference level	Predictors	Estimates [95% CIs]	*SE*	*df*	*t*	*P*
Teacher feedback group & pretest	Intercept	10.00 [6.54 – 13.46]	1.75	7.33	5.71	0.000618**
	Posttest	8.04 [5.38 – 10.70]	1.35	102.29	5.97	3.46E-08***
	Delayed posttest	8.14 [4.64 – 11.64]	1.77	111.26	4.59	1.17E-05***
	AF group	2.94 [-0.93 – 6.81]	1.96	122.58	1.50	0.136
	ATF group	2.22 [-1.66 – 6.11]	1.97	122.31	1.13	0.261
	Posttest: AF	-0.64 [-4.40 – 3.12]	1.91	102.29	-0.34	0.738
	Delayed: AF	1.06 [-4.08 – 6.20]	2.61	112.53	0.41	0.685
	Posttest: ATF	-2.96 [-6.76 – 0.84]	1.92	102.29	-1.54	0.128
	Delayed: ATF	5.53 [0.38 – 10.68]	2.61	112.69	2.12	0.036
Teacher feedback group & posttest	Intercept	18.04 [14.58 – 21.50]	1.75	7.33	10.30	1.28E-05***
	Pretest	-8.04 [-10.70 – -5.38]	1.35	102.29	-5.97	3.46E-08***
	Delayed posttest	0.10 [-3.40 – 3.60]	1.77	111.26	0.06	0.956
	AF group	2.30 [-1.57 – 6.17]	1.96	122.58	1.17	0.243
	ATF group	-0.73 [-4.62 – 3.1]	1.97	122.31	-0.37	0.709
	Pretest: AF	0.64 [-3.12 – 4.40]	1.90	102.29	0.34	0.738
	Delayed: AF	1.70 [-3.44 – 6.84]	2.61	112.53	0.65	0.516
	Pretest: ATF	2.96 [-0.84 – 6.76]	1.92	102.29	1.54	0.128
	Delayed: ATF	8.49 [3.34 – 13.64]	2.61	112.69	3.26	0.002**
Teacher feedback group & delayed test	Intercept	18.14 [14.00 – 22.28]	2.10	14.65	8.65	3.88E-07***
	Pretest	-8.14 [-11.64 – -4.64]	1.77	111.26	-4.59	1.17E-05***
	Posttest	-0.10 [-3.60 – 3.40]	1.77	111.26	-0.06	0.956
	AF group	4.00 [-1.23 – 9.23]	2.65	169.91	1.51	0.133
	ATF group	7.75 [2.54 – 12.96]	2.64	168.35	2.94	0.004*
	Pretest: AF	-1.06 [-6.20 – 4.08]	2.61	112.53	-0.41	0.685
	Posttest: AF	-1.70 [-6.84 – 3.44]	2.61	112.53	-0.65	0.516
	Pretest: ATF	-5.53 [-10.68 – -0.38]	2.61	112.69	-2.12	0.036
	Posttest: ATF	-8.49 [-13.64 – -3.34]	2.61	112.69	-3.26	0.002**

Note. Alpha values adjusted to correct for four repeated tests. *SE* = standard error; *df* = degree of freedom; *AF* = automated feedback group; *ATF* = automated+teacher feedback group.

*$p < .0125$. **$p < .0025$. ***$p < .00025$

G.11 *Summary of fixed effect predictors of mixed-effects model outcome Vocabulary Use score with Automated feedback group as a reference.*

Reference level	Predictors	Estimates [95% CIs]	*SE*	*df*	*t*	*P*
Automated feedback group & pretest	Intercept	12.94 [9.51 – 16.37]	1.74	7.29	7.45	1.17E-04***
	Posttest	7.40 [4.74 – 10.06]	1.35	102.29	5.49	2.88E-07***
	Delayed posttest	9.20 [5.43 – 12.96]	1.91	113.84	4.82	4.45E-06***
	TF group	-2.94 [-6.81 – 0.93]	1.96	122.58	-1.50	0.136
	ATF group	-0.72 [-4.60 – 3.16]	1.96	122.08	-0.37	0.714
	Posttest:TF	0.64 [-3.12 – 4.40]	1.90	102.29	0.34	0.738
	Delayed:TF	-1.06 [-6.20 – 4.08]	2.61	112.53	-0.41	0.685
	Posttest: ATF	-2.32 [-6.12 – 1.48]	1.92	102.29	-1.20	0.231
	Delayed: ATF	4.47 [-0.86 – 9.81]	2.70	113.50	1.65	0.101
Automated feedback group & posttest	Intercept	20.34 [16.91 – 23.77]	1.74	7.29	11.72	5.45E-06***
	Pretest	-7.4 [-10.06 – -4.74]	1.35	102.29	-5.49	2.88E-07***
	Delayed posttest	1.80 [-1.97 – 5.56]	1.91	113.84	0.94	0.348
	TF group	-2.30 [-6.17 – 1.57]	1.96	122.58	-1.17	0.243
	ATF group	-3.04 [-6.91 – 0.84]	1.96	122.08	-1.55	0.125
	Pretest: TF	-0.64 [-4.40 – 3.12]	1.91	102.29	-0.34	0.738
	Delayed: TF	-1.7 [-6.84 – 3.44]	2.61	112.53	-0.65	0.516
	Pretest: ATF	2.32 [-1.48 – 6.12]	1.92	102.29	1.20	0.231
	Delayed: ATF	6.79 [1.45 – 12.13]	2.70	113.50	2.51	0.013*
Automated feedback group & delayed test	Intercept	22.14 [17.79 – 26.48]	2.20	18.07	10.06	7.84E-09***
	Pretest	-9.2 [-12.96 – -5.43]	1.91	113.84	-4.82	4.45E-06***
	Posttest	-1.8 [-5.56 – 1.97]	1.91	113.84	-0.94	0.348
	TF group	-4.00 [-9.23 – 1.23]	2.65	169.91	-1.51	0.133
	ATF group	3.75 [-1.65 – 9.15]	2.73	169.50	1.37	0.172
	Pretest: TF	1.06 [-4.08 – 6.20]	2.61	112.53	0.41	0.685
	Posttest:TF	1.7 [-3.44 – 6.84]	2.61	112.53	0.65	0.516
	Pretest: ATF	-4.47 [-9.81 – 0.86]	2.70	113.50	-1.65	0.101
	Posttest: ATF	-6.79 [-12.13 – -1.45]	2.70	113.50	-2.51	0.013*

Note. Alpha values adjusted to correct for four repeated tests. *SE* = standard error; *df* = degree of freedom; *TF* = teacher feedback group; *ATF* = automated+teacher feedback group.

*$p < .0125$. **$p < .0025$. ***$p < .00025$

G.12 *Summary of fixed effect predictors of mixed-effects model outcome for Vocabulary Use score with Automated+teacher feedback Group as a reference.*

Reference level	Predictors	Estimates [95% CIs]	*SE*	*df*	*t*	*P*
Automated + teacher feedback group & pretest	Intercept	12.22 [8.75 – 15.69]	1.76	7.66	6.95	1.46E-04***
	Posttest	5.08 [2.37 – 7.80]	1.37	102.29	3.70	3.52E-04***
	Delayed posttest	13.67 [9.89 – 17.45]	1.91	113.31	7.14	9.57E-11***
	TF group	-2.22 [-6.11 – 1.66]	1.97	122.31	-1.13	0.261
	AF group	0.72 [-3.16 – 4.60]	1.96	122.08	0.37	0.714
	Posttest: TF	2.96 [-0.84 – 6.76]	1.92	102.29	1.54	0.128
	Delayed: TF	-5.53 [-10.68 – -0.38]	2.61	112.69	-2.12	0.036
	Posttest: AF	2.32 [-1.48 – 6.12]	1.92	102.29	1.20	0.231
	Delayed: AF	-4.47 [-9.81 – 0.86]	2.70	113.50	-1.65	0.101
Automated + teacher feedback group & posttest	Intercept	17.3 [13.83 – 20.78]	1.76	7.66	9.84	1.30E-05***
	Pretest	-5.08 [-7.80 – -2.37]	1.37	102.29	-3.70	3.52E-04***
	Delayed posttest	8.59 [4.81 – 12.37]	1.91	113.31	4.49	1.76E-05***
	TF group	0.73 [-3.15 – 4.62]	1.97	122.31	0.37	0.709
	AF group	3.04 [-0.84 – 6.91]	1.96	122.08	1.55	0.125
	Pretest: TF	-2.96 [-6.76 – 0.84]	1.92	102.29	-1.54	0.128
	Delayed: TF	-8.49 [-13.64 – -3.34]	2.61	112.69	-3.26	0.001**
	Pretest: AF	-2.32 [-6.12 – 1.48]	1.92	102.29	-1.20	0.231
	Delayed: AF	-6.79 [-12.13 – -1.45]	2.70	113.50	-2.51	0.013*
Automated + teacher feedback group & delayed test	Intercept	25.89 [21.53 – 30.25]	2.21	18.27	11.73	6.11E-10***
	Pretest	-13.67 [-17.45 – -9.89]	1.92	113.31	-7.14	9.57E-11***
	Posttest	-8.59 [-12.37 – -4.81]	1.92	113.31	-4.49	1.76E-05***
	TF group	-7.75 [-12.96 – -2.54]	2.64	168.35	-2.94	0.004*
	AF group	-3.75 [-9.15 – 1.65]	2.73	169.50	-1.37	0.172
	Pretest: TF	5.53 [0.38 – 10.68]	2.61	112.69	2.12	0.036
	Posttest: TF	8.49 [3.34 – 13.64]	2.61	112.69	3.26	0.002**
	Pretest: AF	4.47 [-0.86 – 9.81]	2.70	113.50	1.65	0.101
	Posttest: AF	6.79 [1.45 – 12.13]	2.70	113.50	2.51	0.013*

Note. Alpha values adjusted to correct for four repeated tests. *SE* = standard error; *df* = degree of freedom; *TF* = teacher feedback group; *ATF* = automated+teacher feedback group.

*$p < .0125$. **$p < .0025$. ***$p < .00025$

G.13 *Summary of fixed effect predictors of mixed-effects model outcome for Language Use score with Teacher feedback group as a reference.*

Reference level	Predictors	Estimates [95% CIs]	*SE*	*df*	*t*	*P*
Teacher feedback group & pretest	Intercept	9.89 [6.30 – 13.49]	1.82	10.97	5.44	0.00020***
	Posttest	8.8 [5.77 – 11.83]	1.53	101.80	5.74	9.80E-08***
	Delayed posttest	8.93 [4.95 – 12.92]	2.02	110.71	4.43	2.26E-05***
	AF group	3.73 [-0.72 – 8.17]	2.25	121.76	1.66	0.100
	ATF group	1.88 [-2.59 – 6.34]	2.26	121.15	0.83	0.408
	Posttest: AF	-1.12 [-5.40 – 3.16]	2.17	101.80	-0.52	0.607
	Delayed: AF	0.79 [-5.06 – 6.65]	2.97	111.89	0.27	0.790
	Posttest: ATF	-2.93 [-7.25 – 1.40]	2.19	101.80	-1.34	0.185
	Delayed: ATF	5.58 [-0.28 – 11.44]	2.97	111.97	1.88	0.063
Teacher feedback group & posttest	Intercept	18.69 [15.10 – 22.29]	1.82	10.97	10.28	5.78E-07***
	Pretest	-8.8 [-11.83 – -5.77]	1.53	101.80	-5.74	9.80E-08***
	Delayed posttest	0.13 [-3.85 – 4.12]	2.02	110.71	0.07	0.947
	AF group	2.61 [-1.84 – 7.05]	2.25	121.76	1.16	0.249
	ATF group	-1.05 [-5.51 – 3.42]	2.26	121.15	-0.46	0.645
	Pretest: AF	1.12 [-3.16 – 5.40]	2.17	101.80	0.52	0.607
	Delayed: AF	1.91 [-3.94 – 7.77]	2.97	111.89	0.65	0.520
	Pretest: ATF	2.93 [-1.40 – 7.25]	2.19	101.80	1.34	0.185
	Delayed: ATF	8.51 [2.64 – 14.37]	2.97	111.97	2.86	0.005*
Teacher feedback group & delayed test	Intercept	18.83 [14.40 – 23.26]	2.24	23.90	8.39	1.39E-08***
	Pretest	-8.93 [-12.92 – -4.95]	2.02	110.71	-4.43	2.26E-05***
	Posttest	-0.13 [-4.12 – 3.85]	2.02	110.71	-0.07	0.947
	AF group	4.52 [-1.47 – 10.50]	3.03	170.07	1.49	0.138
	ATF group	7.46 [1.50 – 13.43]	3.02	168.40	2.47	0.015
	Pretest: AF	-0.79 [-6.65 – 5.06]	2.97	111.89	-0.27	0.790
	Posttest: AF	-1.91 [-7.77 – 3.94]	2.97	111.89	-0.65	0.520
	Pretest: ATF	-5.58 [-11.44 – 0.28]	2.97	111.97	-1.88	0.063
	Posttest: ATF	-8.51 [-14.37 – -2.64]	2.97	111.97	-2.86	0.005*

Note. Alpha values adjusted to correct for four repeated tests. *SE* = standard error; *df* = degree of freedom; *TF* = teacher feedback group; *ATF* = automated+teacher feedback group.

*$p < .0125$. **$p < .0025$. ***$p < .00025$

G.14 *Summary of fixed effect predictors of mixed-effects model outcome Language Use score with Automated feedback group as a reference.*

Reference level	Predictors	Estimates [95% CIs]	*SE*	*df*	*t*	*P*
Automated feedback group & pretest	Intercept	13.62 [10.06 – 17.18]	1.81	11.31	7.55	9.65E-06***
	Posttest	7.68 [4.65 – 10.71]	1.53	101.80	5.01	2.30E-06***
	Delayed posttest	9.73 [5.44 – 14.02]	2.17	113.08	4.48	1.82E-05***
	TF group	-3.73 [-8.17 – 0.72]	2.25	121.76	-1.66	0.1
	ATF group	-1.85 [-6.30 – 2.61]	2.26	120.70	-0.82	0.415
	Posttest:TF	1.12 [-3.16 – 5.40]	2.17	101.80	0.52	0.607
	Delayed:TF	-0.79 [-6.65 – 5.06]	2.97	111.89	-0.27	0.79
	Posttest: ATF	-1.81 [-6.13 – 2.52]	2.19	101.80	-0.82	0.412
	Delayed: ATF	4.79 [-1.29 – 10.87]	3.08	112.78	1.56	0.123
Automated feedback group & posttest	Intercept	21.3 [17.74 – 24.86]	1.81	11.31	11.80	1.05E-07***
	Pretest	-7.68 [-10.71 – -4.65]	1.53	101.80	-5.01	2.30E-06***
	Delayed posttest	2.05 [-2.24 – 6.34]	2.17	113.08	0.94	0.348
	TF group	-2.61 [-7.05 – 1.84]	2.25	121.76	-1.16	0.249
	ATF group	-3.65 [-8.11 – 0.81]	2.26	120.70	-1.62	0.108
	Pretest: TF	-1.12 [-5.40 – 3.16]	2.17	101.80	-0.52	0.607
	Delayed: TF	-1.91 [-7.77 – 3.94]	2.97	111.89	-0.65	0.520
	Pretest: ATF	1.81 [-2.52 – 6.13]	2.19	101.80	0.82	0.412
	Delayed: ATF	6.59 [0.51 – 12.67]	3.08	112.78	2.14	0.034
Automated feedback group & posttest	Intercept	21.3 [17.74 – 24.86]	1.81	11.31	11.80	1.05E-07***
	Pretest	-7.68 [-10.71 – -4.65]	1.53	101.80	-5.01	2.30E-06***
	Delayed posttest	2.05 [-2.24 – 6.34]	2.17	113.08	0.94	0.348
	TF group	-2.61 [-7.05 – 1.84]	2.25	121.76	-1.16	0.249
	ATF group	-3.65 [-8.11 – 0.81]	2.26	120.70	-1.62	0.108
	Pretest: TF	-1.12 [-5.40 – 3.16]	2.17	101.80	-0.52	0.607
	Delayed: TF	-1.91 [-7.77 – 3.94]	2.97	111.89	-0.65	0.520
	Pretest: ATF	1.81 [-2.52 – 6.13]	2.19	101.80	0.82	0.412
	Delayed: ATF	6.59 [0.51 – 12.67]	3.08	112.78	2.14	0.034

Note. Alpha values adjusted to correct for four repeated tests. *SE* = standard error; *df* = degree of freedom; *TF* = teacher feedback group; *ATF* = automated+teacher feedback group.

*$p < .0125$. **$p < .0025$. ***$p < .00025$

G.15 *Summary of fixed effect predictors of mixed-effects model outcome for Language Use score with Automated+teacher feedback Group as a reference.*

Reference level	Predictors	Estimates [95% CIs]	*SE*	*df*	*t*	*P*
Automated + teacher feedback group & pretest	Intercept	11.77 [8.15 – 15.39]	1.83	12.00	6.42	3.30E-05***
	Posttest	5.88 [2.79 – 8.96]	1.57	101.80	3.76	2.89E-04***
	Delayed posttest	14.52 [10.21 – 18.82]	2.18	112.59	6.66	1.07E-09***
	TF group	-1.88 [-6.34 – 2.59]	2.26	121.15	-0.83	0.408
	AF group	1.85 [-2.61 – 6.30]	2.26	120.70	0.82	0.415
	Posttest: TF	2.92 [-1.40 – 7.25]	2.19	101.80	1.34	0.185
	Delayed: TF	-5.58 [-11.44 – 0.28]	2.97	111.97	-1.88	0.063
	Posttest: AF	1.81 [-2.52 – 6.13]	2.19	101.80	0.82	0.412
	Delayed: AF	-4.79 [-10.87 – 1.29]	3.08	112.78	-1.56	0.123
Automated + teacher feedback group & posttest	Intercept	17.65 [14.03 – 21.27]	1.83	12.00	9.63	5.40E-07***
	Pretest	-5.88 [-8.96 – -2.79]	1.57	101.80	-3.76	2.89E-04***
	Delayed posttest	8.64 [4.34 – 12.95]	2.18	112.59	3.96	0.0001***
	TF group	1.05 [-3.42 – 5.51]	2.26	121.15	0.46	0.645
	AF group	3.65 [-0.81 – 8.11]	2.26	120.70	1.62	0.108
	Pretest: TF	-2.93 [-7.25 – 1.40]	2.19	101.80	-1.34	0.185
	Delayed: TF	-8.51 [-14.37 – -2.64]	2.97	111.97	-2.86	0.005*
	Pretest: AF	-1.81 [-6.13 – 2.52]	2.19	101.80	-0.82	0.412
	Delayed: AF	-6.59 [-12.67 – -0.51]	3.08	112.78	-2.14	0.034
Automated + teacher feedback group & delayed test	Intercept	26.29 [21.59 – 30.99]	2.38	31.28	11.04	2.57E-12***
	Pretest	-14.5 [-18.82 – -10.21]	2.18	112.59	-6.66	1.07E-09***
	Posttest	-8.64 [-12.95 – -4.34]	2.18	112.59	-3.96	0.0001***
	TF group	-7.46 [-13.43 – -1.50]	3.02	168.40	-2.47	0.015
	AF group	-2.94 [-9.12 – 3.24]	3.13	169.71	-0.94	0.349
	Pretest: TF	5.58 [-0.28 – 11.44]	2.97	111.97	1.88	0.063
	Posttest: TF	8.51 [2.64 – 14.37]	2.97	111.97	2.86	0.005*
	Pretest: AF	4.79 [-1.29 – 10.87]	3.08	112.78	1.56	0.123
	Posttest: AF	6.59 [0.51 – 12.67]	3.08	112.78	2.14	0.034

Note. Alpha values adjusted to correct for four repeated tests. *SE* = standard error; *df* = degree of freedom; *TF* = teacher feedback group; *ATF* = automated+teacher feedback group.

*$p < .0125$. **$p < .0025$. ***$p < .00025$

G.16 *Summary of fixed effect predictors of mixed-effects model outcome for Mechanics score with Teacher feedback group as a reference.*

Reference level	Predictors	Estimates [95% CIs]	*SE*	*df*	*t*	*P*
Teacher feedback group & pretest	Intercept	7.70	1.83	8.83	4.20	0.002408**
	Posttest	6.73	1.66	100.99	4.05	0.0001***
	Delayed posttest	7.48	2.16	115.44	3.47	0.000736**
	AF group	2.95	2.05	143.73	1.44	0.152
	ATF group	0.64	2.06	145.14	0.31	0.758
	Posttest: AF	-3.18	2.35	100.75	-1.36	0.178
	Delayed: AF	1.28	3.16	116.69	0.40	0.687
	Posttest: ATF	-1.27	2.38	101.01	-0.53	0.596
	Delayed: ATF	-0.67	3.18	117.52	-0.21	0.834
Teacher feedback group & posttest	Intercept	14.43	1.84	8.82	7.87	2.84E-05***
	Pretest	-6.73	1.66	100.99	-4.05	1.00E-04***
	Delayed posttest	0.74	2.16	115.95	0.34	0.731
	AF group	-0.23	2.05	144.65	-0.11	0.910
	ATF group	-0.63	2.06	144.51	-0.31	0.760
	Pretest: AF	3.18	2.35	100.75	1.36	0.178
	Delayed: AF	4.46	3.16	116.88	1.41	0.161
	Pretest: ATF	1.27	2.38	101.01	0.53	0.596
	Delayed: ATF	0.60	3.17	118.18	0.19	0.851
Teacher feedback group & delayed test	Intercept	15.18	2.30	20.43	6.61	1.74E-06***
	Pretest	-7.48	2.16	115.44	-3.47	7.36E-04***
	Posttest	-0.74	2.16	115.95	-0.34	0.731
	AF group	4.23	2.95	168.95	1.43	0.154
	ATF group	-0.03	2.94	169.32	-0.01	0.992
	Pretest: AF	-1.28	3.16	116.69	-0.40	0.687
	Posttest: AF	-4.46	3.16	116.88	-1.41	0.161
	Pretest: ATF	0.67	3.18	117.52	0.21	0.834
	Posttest: ATF	-0.60	3.17	118.18	-0.19	0.850

Note. Alpha values adjusted to correct for four repeated tests. *SE* = standard error; *df* = degree of freedom; *TF* = teacher feedback group; *ATF* = automated+teacher feedback group.

*$p < .0125$. **$p < .0025$. ***$p < .00025$

G.17 *Summary of fixed effect predictors of mixed-effects model outcome Mechanics score with Automated feedback group as a reference.*

Reference level	Predictors	Estimates [95% CIs]	*SE*	*df*	*t*	*P*
Automated feedback group & pretest	Intercept	10.65	1.82	8.79	5.85	2.67E-04***
	Posttest	3.55	1.67	100.75	2.13	.035
	Delayed posttest	8.75	2.32	120.10	3.78	2.46E-04***
	TF group	-2.95	2.05	143.73	-1.44	0.152
	ATF group	-2.32	2.06	145.06	-1.12	0.263
	Posttest:TF	3.18	2.35	100.75	1.36	0.178
	Delayed:TF	-1.28	3.16	116.69	-0.40	0.687
	Posttest: ATF	1.92	2.40	96.99	0.80	0.426
	Delayed: ATF	-1.94	3.31	114.40	-0.59	0.558
Automated feedback group & posttest	Intercept	14.20	1.83	8.80	7.78	3.14E-05***
	Pretest	-3.55	1.67	100.75	-2.13	.035
	Delayed posttest	5.20	2.34	110.79	2.23	0.0281
	TF group	0.23	2.05	144.65	0.11	0.910
	ATF group	-0.40	2.06	144.77	-0.19	0.847
	Pretest: TF	-3.18	2.35	100.75	-1.36	0.178
	Delayed: TF	-4.46	3.16	116.88	-1.41	0.161
	Pretest: ATF	-1.92	2.40	96.99	-0.80	0.426
	Delayed: ATF	-3.86	3.30	116.30	-1.17	0.245
Automated feedback group & delayed test	Intercept	19.41	2.44	25.60	7.95	2.21E-08***
	Pretest	-8.75	2.32	120.10	-3.78	2.46E-04***
	Posttest	-5.20	2.34	110.79	-2.23	0.028
	TF group	-4.23	2.95	168.95	-1.43	0.154
	ATF group	-4.26	3.07	166.56	-1.39	0.168
	Pretest: TF	1.28	3.16	116.69	0.40	0.687
	Posttest:TF	4.46	3.16	116.88	1.41	0.161
	Pretest: ATF	1.94	3.31	114.40	0.59	0.558
	Posttest: ATF	3.86	3.30	116.30	1.17	0.245

Note. Alpha values adjusted to correct for four repeated tests. *SE* = standard error; *df* = degree of freedom; *TF* = teacher feedback group; *ATF* = automated+teacher feedback group.

*$p < .0125$. **$p < .0025$. ***$p < .00025$

G.18 *Summary of fixed effect predictors of mixed-effects model outcome for Mechanics score with Automated+teacher feedback Group as a reference.*

Reference level	Predictors	Estimates [95% CIs]	*SE*	*df*	*t*	*P*
Automated + teacher feedback group & pretest	Intercept	8.34	1.85	9.24	4.50	.0014**
	Posttest	5.47	1.73	85.59	3.17	.0021**
	Delayed posttest	6.81	2.34	116.17	2.91	.0043*
	TF group	-0.64	2.06	145.14	-0.31	0.758
	AF group	2.31	2.06	145.06	1.12	0.263
	Posttest: TF	1.27	2.38	101.01	0.53	0.596
	Delayed: TF	0.67	3.18	117.52	0.21	0.834
	Posttest: AF	-1.92	2.40	96.99	-0.80	0.426
	Delayed: AF	1.94	3.31	114.40	0.59	0.558
Automated + teacher feedback group & posttest	Intercept	13.80	1.85	9.24	7.46	3.32E-05***
	Pretest	-5.47	1.73	85.59	-3.17	.0021**
	Delayed posttest	1.34	2.32	117.77	0.58	0.563
	TF group	0.63	2.06	144.51	0.31	0.760
	AF group	0.40	2.06	144.77	0.19	0.847
	Pretest: TF	-1.27	2.38	101.01	-0.53	0.596
	Delayed: TF	-0.60	3.17	118.18	-0.19	0.850
	Pretest: AF	1.92	2.40	96.99	0.80	0.426
	Delayed: AF	3.86	3.30	116.30	1.17	0.245
Automated + teacher feedback group & delayed test	Intercept	15.15	2.43	26.01	6.22	1.39E-06***
	Pretest	-6.81	2.34	116.17	-2.91	.0043*
	Posttest	-1.34	2.32	117.77	-0.58	0.563
	TF group	0.03	2.94	169.32	0.01	0.992
	AF group	4.26	3.07	166.56	1.39	0.168
	Pretest: TF	-0.67	3.18	117.52	-0.21	0.834
	Posttest: TF	0.60	3.17	118.18	0.19	0.851
	Pretest: AF	-1.94	3.31	114.40	-0.59	0.558
	Posttest: AF	-3.86	3.30	116.30	-1.17	0.245

Note. Alpha values adjusted to correct for four repeated tests. *SE* = standard error; *df* = degree of freedom; *TF* = teacher feedback group; *ATF* = automated+teacher feedback group.

*$p < .0125$. **$p < .0025$. ***$p < .00025$

G.19 *Summary of fixed effect predictors of mixed-effects model outcome for total writing production data of teacher feedback group with an interaction between test time and genre: (Score~ Test*Genre + (1+Test*Genre|Teacher)+ (Test*Genre|Version) + (1+Test*Genre|Subject)*

Reference level	Predictors	Estimates [95% CIs]	*SE*	*df*	*t*	*P*
Pretest & expository genre	(Intercept)	30.24 [19.39 – 41.09]	5.52	3.23	5.47	0.0099*
	Posttest	12.84 [5.18 – 20.50]	3.90	25.78	3.29	0.0029**
	Delayed	16.39 [5.12 – 27.67]	5.74	16.43	2.86	0.011*
	GenrePersuasive	-11.52 [-16.29 – -6.75]	2.43	29.19	-4.74	5.2E-05***
	Posttest:Persuasive	12.04 [9.16 – 14.92]	1.47	530.39	8.22	1.6E-15***
	Delayed:Persuasive	10.52 [6.68 – 14.37]	1.96	544.98	5.38	1.1E-07***
Posttest & expository genre	(Intercept)	43.08 [32.35 – 53.81]	5.46	3.47	7.89	0.00247**
	Pretest	-12.84 [-20.50 – -5.18]	3.90	25.79	-3.29	0.00289**
	Delayed	3.55 [-3.94 – 11.05]	3.82	7.76	0.93	0.3797
	GenrePersuasive	0.52 [-4.25 – 5.29]	2.43	29.20	0.21	.831
	Pretest: Persuasive	-12.04 [-14.92 – -9.16]	1.47	530.39	-8.22	1.6E-15***
	Delayed: Persuasive	-1.52 [-5.36 – 2.33]	1.96	544.98	-0.78	.439
Delayed posttest & expository genre	(Intercept)	46.63 [32.11 – 61.16]	7.39	8.78	6.31	0.00015***
	Pretest	-16.39 [-27.67 – -5.12]	5.74	16.43	-2.86	0.0112*
	Posttest	-3.55 -11.05 – 3.94]	3.82	7.76	-0.93	0.379
	Persuasive	-1.00 [-6.41 – 4.42]	2.76	46.83	-0.36	.719
	Pretest:Persuasive	-10.52 [-14.37 – -6.68]	1.96	544.98	-5.38	1.1E-07***
	Posttest:Persuasive	1.52 [-2.33 – 5.36]	1.96	544.98	0.78	.439
Pretest & persuasive genre	(Intercept)	18.72 [7.99 – 29.45]	5.47	2.98	3.43	0.042
	Posttest	24.88 [17.22 – 32.54]	3.90	25.79	6.38	9.7E-07***
	Delayed	26.92 [15.86 – 37.97]	5.63	16.26	4.78	0.00019***
	Expository	11.52 [6.75 – 16.29]	2.43	29.19	4.74	5.2E-05***
	Posttest:Expository	-12.04 [-14.92 – -9.16]	1.47	530.39	-8.22	1.6E-15***
	Delayed:Expository	-10.52 [-14.37 – -6.68]	1.96	544.98	-5.38	1.1E-07***
Posttest & persuasive genre	(Intercept)	43.6 [33.61 – 53.59]	5.088	2.55	8.57	0.006*
	Pretest	-24.88 [-32.54 – -17.22]	3.901	25.777	-6.38	9.7E-07***
	Delayed	2.038 [-5.13 – 9.20]	3.649	7.517	0.56	0.592
	Expository	-0.52 [-5.29 – 4.25]	2.431	29.192	-0.21	0.832
	Pretest: Expository	12.04 [9.16 – 14.92]	1.465	530.4	8.22	1.6E-15***
	Delayed: Expository	1.52 [-2.33 – 5.36]	1.956	544.98	0.78	0.438

Note. Alpha values adjusted to correct for two repeated tests. *$p < .025$. **$p < .005$. ***$p < .0005$

SE = standard error; *df* = degree of freedom.

G.20 *Summary of fixed effect predictors of mixed-effects model outcome for total writing production data of automated feedback group with an interaction between test time and genre: (Score~ Test*Genre + (1+Test*Genre|Teacher)+ (Test*Genre|Version) + (1+Test*Genre|Subject)*

Reference level	Predictors	Estimates [95% CIs]	*SE*	*df*	*t*	*P*
Pretest & expository genre	(Intercept)	35.8 [27.92 – 43.68]	4.01	24.16	8.92	4.0E-09***
	Posttest	14.52 [8.98 – 20.06]	2.82	25.09	5.15	2.5E-05***
	Delayed	23.76 [12.63 – 34.88]	5.67	8.02	4.19	0.0030**
	GenrePersuasive	-7.52 [-12.57 – -2.47]	2.57	25.47	-2.92	7.2E-03*
	Posttest:Persuasive	4.52 [2.87 – 6.17]	0.84	512.95	5.40	1.1E-07***
	Delayed:Persuasive	9.00 [6.61 – 11.40]	1.22	517.74	7.39	5.9E-13***
Posttest & expository genre	(Intercept)	50.32 [42.89 – 57.75]	3.78	24.17	13.3	1.3E-12***
	Pretest	-14.52 [-20.06 – -8.98]	2.82	25.10	-5.15	2.5E-05***
	Delayed	9.26 [-0.44 – 18.96]	4.94	9.74	1.88	0.09098
	GenrePersuasive	-3 [-8.06 – 2.06]	2.57	25.45	-1.17	.26
	Pretest: Persuasive	-4.52 [-6.17 – -2.87]	0.84	512.97	-5.40	1.1E-07***
	Delayed: Persuasive	4.48 [2.09 – 6.88]	1.22	517.76	3.68	2.6E-04***
Delayed posttest & expository genre	(Intercept)	59.58 [48.22 – 70.94]	5.78	9.76	10.3	1.5E-06***
	Pretest	-23.78 [-34.93 – -12.63]	5.68	7.98	-4.19	3.06E-03**
	Posttest	-9.26 [-18.95 – 0.44]	4.94	9.75	-1.88	0.091
	Persuasive	1.48 [-3.86 – 6.83]	2.72	31.70	0.55	.590
	Pretest:Persuasive	-9.00 [-11.40 – -6.61]	1.22	517.76	-7.39	5.9E-13***
	Posttest:Persuasive	-4.48 [-6.88 – -2.09]	1.22	517.76	-3.68	2.6E-04***
Pretest & persuasive genre	(Intercept)	28.28 [19.86 – 36.70]	4.29	24.13	6.60	7.8E-07***
	Posttest	19.04 [13.50 – 24.58]	2.82	25.08	6.75	4.4E-07***
	Delayed	32.79 [21.56 – 44.02]	5.72	8.01	5.74	0.00043***
	Expository	7.52 [2.46 – 12.58]	2.57	25.45	2.92	.0072*
	Posttest:Expository	-4.52 [-6.17 – -2.87]	0.84	512.97	-5.40	1.1E-07***
	Delayed:Expository	-9.00 [-11.40 – -6.61]	1.22	517.75	-7.39	5.9E-13***
Posttest & persuasive genre	(Intercept)	47.32 [39.32 – 55.32]	4.07	24.14	11.6	2.3E-11***
	Pretest	-19.04 [-24.58 – -13.50]	2.82	25.09	-6.76	4.4E-07***
	Delayed	13.743 {3.96 – 23.53]	4.98	9.82	2.76	0.020*
	Expository	3 [-2.06 – 8.06]	2.57	25.44	1.17	0.254
	Pretest: Expository	4.52 [2.87 – 6.17]	0.84	512.97	5.40	1.1E-07***
	Delayed: Expository	-4.48 [-6.88 – -2.09]	1.22	517.76	-3.68	0.00026***

Note. Alpha values adjusted to correct for two repeated tests. *$p < .025$. **$p < .005$. ***$p < .0005$

SE = standard error; *df* = degree of freedom.

G.21 *Summary of fixed effect predictors of mixed-effects model outcome for total writing production data of automated+teacher feedback group with an interaction between test time and genre: (Score~ Test*Genre + (1+Test*Genre|Teacher)+ (Test*Genre|Version) + (1+Test*Genre|Subject)*

Reference level	Predictors	Estimates [95% CIs]	*SE*	*df*	*t*	*P*
Pretest & expository genre	(Intercept)	32.87 [26.23 – 39.52]	3.38	23.37	9.72	1.1E-09***
	Posttest	9.37 [1.81 – 16.94]	3.85	24.02	2.43	0.023*
	Delayed	24.25 [17.08 – 31.42]	3.65	15.39	6.64	6.9E-06***
	GenrePersuasive	-8.33 [-13.03 – -3.64]	2.39	26.44	-3.49	0.00173**
	Posttest:Persuasive	8.17 [5.95 – 10.39]	1.13	496.21	7.23	1.9E-12***
	Delayed:Persuasive	10.71 [7.55 – 13.87]	1.61	505.36	6.65	7.4E-11***
Posttest & expository genre	(Intercept)	42.25 [35.87 – 48.63]	3.25	23.38	13.0	3.4E-12***
	Pretest	-9.38 [-16.94 – -1.81]	3.85	24.02	-2.43	0.023*
	Delayed	14.87 [8.10 – 21.65]	3.45	11.40	4.31	0.0011**
	GenrePersuasive	-0.17 [-4.86 – 4.53]	2.39	26.44	-0.07	0.944
	Pretest: Persuasive	-8.17 [-10.39 – -5.95]	1.13	496.21	-7.23	1.9E-12***
	Delayed: Persuasive	2.54 [-0.62 – 5.70]	1.61	505.36	1.58	0.11
Delayed posttest & expository genre	(Intercept)	57.12 [51.69 – 62.55]	2.77	13.55	20.7	1.2E-11***
	Pretest	-24.25 [-31.42 – -17.08]	3.65	15.38	-6.64	6.9E-06***
	Posttest	-14.87 [-21.65 – -8.10]	3.45	11.40	-4.31	0.0011**
	Persuasive	2.37 [-2.83 – 7.58]	2.65	39.44	0.90	0.376
	Pretest:Persuasive	-10.71 [-13.87 – -7.55]	1.61	505.36	-6.65	7.4E-11***
	Posttest:Persuasive	-2.54 [-5.70 – 0.62]	1.61	505.36	-1.58	0.115
Pretest & persuasive genre	(Intercept)	24.54 [18.00 – 31.09]	3.33	23.40	7.37	1.6E-07***
	Posttest	17.54 [9.97 – 25.11]	3.85	24.02	4.55	1.3E-04***
	Delayed	34.96 [27.78 – 42.13]	3.65	15.32	9.57	7.4E-08***
	Expository	8.33 [3.64 – 13.03]	2.39	26.44	3.49	1.7E-03**
	Posttest:Expository	-8.17 [-10.39 – -5.95]	1.13	496.21	-7.23	1.9E-12***
	Delayed:Expository	-10.70 [-13.87 – -7.55]	1.61	505.36	-6.65	7.4E-11***
Posttest & persuasive genre	(Intercept)	42.08 [35.14 – 49.02]	3.53	23.33	11.9	2.1E-11***
	Pretest	-17.54 [-25.11 – -9.97]	3.85	24.02	-4.55	1.3E-04***
	Delayed	17.4154 [10.64 – 24.19]	3.45	11.39	5.05	0.00034***
	Expository	0.17 [-4.53 – 4.86]	2.39	26.43	0.07	0.94
	Pretest: Expository	8.17 [5.95 – 10.39]	1.13	496.22	7.23	1.9E-12***
	Delayed: Expository	-2.54 [-5.70 – 0.62]	1.61	505.36	-1.58	0.114

Note. Alpha values adjusted to correct for two repeated tests. *$p < .025$. **$p < .005$. ***$p < .0005$.

SE = standard error; *df* = degree of freedom.

Appendix H: Consent form

Department of Education
Heslington, York, YO10 5DD
Tel: (01904) 323460
Web: *http://www.york.ac.uk/education*

Exploring the Impact of Automated Feedback on Students Performance in Writing Assessment

Consent Form
استمارة الموافقة للمشاركة بالدراسة البحثية

Please tick each box if you are happy to take part in this research.

الرجاء وضع علامة صح في كل المربعات ادناه إن كنت راغبا في المشاركة في الدراسة.

Statement of consent	Tick each box
I confirm that I have read and understood the information given to me about the above-named research project and I understand that this will involve me taking part as described above. أؤكد أنني قد قرأت وفهمت المعلومات المعطاة لي عن المشروع البحثي المذكور أعلاه وأنا على وعي بأن ذلك يعني أني سأشارك بالبحث كما هو موضح أعلاه	
I understand that participation in this study is voluntary and that if I wish to withdraw, I can do so at any time during data collection and up to two weeks after the end of data collection (i.e. two weeks after the delayed writing task) أقر بأنني على وعي بأنه يمكنني الانسحاب من الدراسة في أي وقت خلال الدراسة الى مدة أقصاها أسبوعين من تاريخ نهاية جمع البيانات	
I understand that my data will not be identifiable and the anonymous data may be used in publications, presentations and online. أقر بأنني على وعي بأن البيانات سرية و قد تستخدم مجهولة الهوية لأغراض بحثية.	
I am happy to pass the ownership of my data to the researcher. لا مانع من استخدام الباحث للبيانات التي سأشارك بها	
I confirm that I have read the information about GDPR أقر أنني قرأت المعلومات المرفقة عن نظام حماية البيانات المتبع في جامعة يورك	

Name: _____

Signature: _____

Date: _____

Department of Education

Heslington, York, YO10 5DD

Tel: (01904) 323460

Web: *http://www.york.ac.uk/education*

Information Page

صفحة البيانات

Exploring the Impact of Automated Feedback on Students Performance in Writing Assessment

دراسة تأث ري التغذية الراجعة الألية على أداء الطالبات ˙ يف اختبار الكتابة

Dear Student,

I, Nadia Junaid, am currently carrying out a research project to explore the impact of automated feedback on students' performance in writing assessment. I would like to invite you to take part in this research project.

Before agreeing to take part, please read this information sheet carefully and let me know if anything is unclear or you would like further information.

عزيز : يت الطالبة

اعمل حاليا على تنفيذ م رشوع بح $ يث حول تأث ري التغذية الراجعة الألية على أداء الطالبات ˙ يف اختبار الكتابة، و الهدف من هذا الخطاب دعوتكم للمشاركة ˙ يف هذه الدراسة.

قبل الموافقة على المشاركة، يربج قراءة ورقة المعلومات هذه بعناية وإعلا يم إذا كان أي ر يشء غ ري واضح أو إذا كنت ترغ ى يث ˙ يف الحصول على مزيد من المعلومات.

Purpose of the study

The study is designed to investigate to what extent the automated feedback on students' writing is effective. To answer this question a quasi-experimental study is going to be employed to compare between three conditions of providing feedback on writing: teacher feedback only, automated feedback only and a combination of teacher and automated feedback.

الغرض من الدراسة

تم تصميم الدراسة للتحقق من مدى فعالية التغذية الراجعة الآلية على مستوى الكتابة لدى الطالبات. للإجابة على هذا السؤال، سيتم استخدام دراسة شبه تجريبية مقارنة للمقارنة ب ر ·ي ثلاثة طرق لتقديم التغذية الراجعة حول الكتابة: التغذية الراجعة من المعلم فقط، التغذية الراجعة الآلية فقط، والجمع ب ر ·ي التغذية الراجعة من المعلم والتغذية الراجعة الآلية.

What would this mean for you?

You will be invited to join one of the three groups: teacher feedback, automated feedback, or teacher+automated feedback group. The intervention will last for four weeks. During the intervention you will take part in writing essays of two different genres. All participants will be invited to complete: a pre-writing task, immediate post writing task, and delayed post writing task.

ماذا يع ˙ ين هذا بالنسبة لك؟

ستتم دعوتك للانضمام إلى إحدى المجموعات الثلاث. ستستمر الدراسة لمدة أربعة أسابيع. أثناء ذلك ستشارك ˙ يف كتابة مقالات تحت نوع ر ·ي مختلف ر ·ي. علاوة على ذلك، سيتم دعوة جميع المشارك ر ·ي لإكمال ثلاث اختبارات: اختبار مسبق واختبار لاحق فوري واختبار لاحق متأخر.

Participation is voluntary

Participation is optional. If you do decide to take part, you will be given a copy of this information sheet for your records and will be asked to complete a consent form. If you change your mind at any point during the study and up to **two weeks** after the end of data collection, you will be able to withdraw your participation without having to provide a reason.

المشاركة اختيارية. إذا قررت المشاركة، فسيتم إعطاؤك نسخة من ورقة المعلومات هذه وسيُطلب منك إكمال نموذج الموافقة.

، فستتمكن من سحب مشاركتك دون إذا غ ريت رأيك ُ يف أي وقت أثناء الدراسة الى مده أقصاها أسبوع ر ُي بعد نهاية ف ُية جمع البيانات

الحاجة إلى إبداء سبب.

Anonymity and confidentiality

The data that you provide (e.g. writing task scores) will be stored by a code number in a password protected computer, and only the researcher will have access to it. Any information that identifies you will be stored separately from the data.

سرية البيانات

عزيز ُ يت الطالبة: سوف يتم التعامل مع جميع البيانات ال ُ يث يتم جمعها من كل المشارك ر ُي ُ يف هذه الدراسة بشية تامة ، و لن يكون بمقدور احد- سوى الباحث- الاطلاع على هذه البيانات.

Storing and using data

Paper-based writing tasks will be stored in secure filing cabinets. It will be fully anonymised two weeks after the delayed written task. Anonymised data will be kept for 10 years in line with the University of York Research Data Management Policy, after which time it will be destroyed. Data may also be used in different ways, e.g., publications and presentations for research or instruction purposes. Please indicate on the consent form enclosed with a tick ☑ if you are happy for this anonymised data to be used in the ways listed.

تخزين واستخدام البيانات

سيتم تخزين البيانات ˙ يف خزائن حفظ آمنة وعلى جهازكمبيوتر الباحثة المح يم بكلمة مرور. سيتم إخفاء هوية البيانات بالكامل بعد أسبوع ر ·ي من انتهاء التجربة.

سيتم الاحتفاظ بالبيانات مجهولة المصدر لمدة ع رش سنوات و قد يتم استخدام البيانات لأغراض بحثية مثل الن رش العل يم والندوات.

Questions or concerns

If you have any questions about this participant information sheet or concerns about how your data is being processed, please feel free to contact **the researcher Nadia Junaid** by email naj515@york.ac.uk , or the Chair of Ethics Committee via email education-research-admin@york.ac.uk. If you are still dissatisfied, please contact the University of York Data Protection Officer at dataprotection@york.ac.uk.

عزيز ٔيت الطالبة: نأمل منك المشاركة ٔيف هذه الدراسة المهمة. إن كان لديك أسئلة تتعلق بهذه الدراسة، فإنه يسعدنا أن تتواص يلى م يع على

ال بيد naj515@york.ac.uk أو على بريد رئيس لجنة أخلاقيات البحث بجامعة يورك

education-research-admin@york.ac.uk

I hope that you will agree to take part. If you are happy to participate, please complete the form enclosed and return it to your class teacher.

عزيزتي الطالبة: إن كنت راغبة بالمشاركة في الدراسة ، فنرجو منك الإجابة على الإستمارة الملحقة.

Please keep this information sheet for your own records.

Thank you for taking the time to read this information.

Yours sincerely,

Nadia Junaid

شاكرة حسن تعاونكم و أرجو منك الاحتفاظ باستمارة المعلومات هذه

مع خالص التحية والتقدير نادية

أحمد جنيد

Appendix I: Information about GDPR

Processing personal data

Under the General Data Protection Regulation (GDPR), the University has to identify a legal basis for processing personal data and, where appropriate, an additional condition for processing special category data.

In line with our charter which states that we advance learning and knowledge by teaching and research, the University processes personal data for research purposes under Article 6 (1)(e) of the GDPR:
Processing is necessary for the performance of a task carried out in the public interest

Special category data is processed under Article 9 (2) (j):

Processing is necessary for archiving purposes in the public interest, or scientific and historical research purposes or statistical purposes

Research will only be undertaken where ethical approval has been obtained, where there is a clear public interest and where appropriate safeguards have been put in place to protect data.

In line with ethical expectations and in order to comply with common law duty of confidentiality, we will seek your consent to participate where appropriate. This consent will not, however, be our legal basis for processing your data under the GDPR.

Protecting and storing personal data

Information that research participants provide will be treated confidentially and shared on a need-to-know basis only. The University is committed to the principle of data protection by design and default and will collect the minimum amount of data necessary for the project. In addition, we will anonymise or pseudonymise data wherever possible.

We will put in place appropriate technical and organisational measures to protect your personal data and/or special category data (for example, data may be stored in secure filing cabinets and/or on a password protected computer).

Sharing of data

The default position is that personal data will only be accessible to members of the project team. In some cases, however, the research may be of a collaborative nature and hence the data will be made accessible to others from outside the University. Information specific to the project will include details of when this is the case, who the 3rd parties are, and what they will do with the data. It is possible that personal data may be shared anonymously with others for secondary research and/or teaching purposes.

Transfer of data internationally

The default position is that data will be stored on University devices and held within the European Economic Area in full compliance with data protection legislation.

However, data may be transferred to the project partners based outside the European Economic Area. Any international transfer will be undertaken in full compliance with the GDPR.

The University has access to cloud storage provided by Google which means that data can be located at any of Google's globally spread data centres. The University has data protection compliant arrangements in place with this provider. For further information see,

https://www.york.ac.uk/it-services/google/policy/privacy/

Your rights in relation to your data

Under the GDPR, you have a general right of access to your data, a right to rectification, erasure, restriction, objection or portability. You also have a right to withdrawal. Please note, not all rights apply where data is processed purely for research purposes. For information see, https://www.york.ac.uk/records-management/generaldataprotectionregulation/individualrights/

Right to complain

If you are unhappy with the way in which your personal data has been handled, you have a right to complain to the Information Commissioner's Office. For information on reporting a concern to the Information Commissioner's Office, see www.ico.org.uk/concerns

References

Al-Hejin, B. (2004). Attention and awareness: Evidence from cognitive and second language acquisition research. *Studies in Applied Linguistics and TESOL*, 4(1). DOI:10.7916/salt.v4i1.1600

Aljaafreh, A., & Lantolf, J. P. (1994). Negative Feedback as Regulation and Second Language Learning in the Zone of Proximal Development. *The Modern Language Journal*, *78*(4), 465–483. https://about.jstor.org/terms

Attali, Y. (2004). Exploring the Feedback and Revision Features of Criterion. *Journal of Second Language Writing*, *14*(3), 191–205. www.ets.org/legal/copyright

Attali, Y., & Burstein, J. (2006). Automated Essay Scoring With e-rater ® V.2. In *The Journal of Technology, Learning, and Assessment* (Vol. 4, Issue 3). www.jtla.org

Azmi, A. M., Al-Jouie, M. F., & Hussain, M. (2019). AAEE – Automated evaluation of students' essays in Arabic language. *Information Processing and Management*, *56*(5), 1736–1752. https://doi.org/10.1016/j.ipm.2019.05.008

Barrot, J. S. (2023). Trends in automated writing evaluation systems research for teaching, learning, and assessment: A bibliometric analysis. *Education and Information Technologies*, *29*(6), 7155–7179. https://doi.org/10.1007/s10639-023-12083-y

Bax, S. (2003). CALL-past, present and future. *System*, *31*, 13–28. www.elsevier.com/locate/system

Bitchener, J. (2008). Evidence in support of written corrective feedback. *Journal of Second Language Writing*, *17*(2), 102–118. https://doi.org/10.1016/j.jslw.2007.11.004

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *International Journal of Phytoremediation*, *21*(1), 7–74. https://doi.org/10.1080/0969595980050102

Bornstein, M. H., & Bruner, J. S. (2014). *Interaction in human development.* . Psychology Press.

Bridgeman, B. (2013). Human ratings and automated essay evaluation. . In M. D. Shermis & J. Burstein (Eds.), *In Handbook of automated essay evaluation* (pp. 221–232). Routledge.

Burstein, J., Chodorow, M., & Leacock, C. (2003). Criterion SM Online Essay Evaluation: An Application for Automated Evaluation of Student Essays. *IAAI*. www.toefl.org/educator/edtwegui.html

Butler, D. L., & Winne, P. H. (1995). Feedback and Self-Regulated Learning: A Theoretical Synthesis. *Review of Educational Research*, *65*(3), 245–281.

Chandler, J. (2003). The efficacy of various kinds of error feedback for improvement in the accuracy and fluency of L2 student writing. *Journal of Second Language Writing*, *12*(3), 267–296. https://doi.org/10.1016/S1060-3743(03)00038-9

Chapelle, C. A. (2017). Evaluation of Technology and Language Learning. In *The Handbook of Technology and Second Language Teaching and Learning* (pp. 378–392). Wiley. https://doi.org/10.1002/9781118914069.ch25

Chapelle, C. A., & Sauro, S. (2017). Introduction to the *Handbook of Technology and Second Language Teaching and Learning*. In *The Handbook of Technology and Second Language Teaching and Learning* (pp. 1–9). Wiley. https://doi.org/10.1002/9781118914069.ch1

Cheng, G. (2017). The impact of online automated feedback on students' reflective journal writing in an EFL course. *Internet and Higher Education*, *34*, 18–27. https://doi.org/10.1016/j.iheduc.2017.04.002

Cheng, G., Law, E., & Wong, T. L. (2017). Investigating effects of automated feedback on EFL students' reflective learning skills. *Proceedings of 2016 IEEE International Conference on Teaching, Assessment and Learning for Engineering, TALE 2016*, 226–230. https://doi.org/10.1109/TALE.2016.7851798

Cohen, Y., Levi, E., & Ben-Simon, A. (2018). Validating human and automated scoring of essays against "True" scores. *Applied Measurement in Education*, *31*(3), 241–250. https://doi.org/10.1080/08957347.2018.1464450

Condon, W. (2013). Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings? *Assessing Writing*, *18*(1), 100–108. https://doi.org/10.1016/j.asw.2012.11.001

Crossley, S. A. (2013). Advancing research in second language writing through computational tools and machine learning techniques: A research agenda. *Language Teaching*, *46*(2), 256–271. https://doi.org/10.1017/S0261444812000547

Deane, P. (2013). Covering the Construct An Approach to Automated Essay Scoring Motivated by a Socio-Cognitive Framework for Defining Literacy Skills. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation : Current applications and new directions*. Routledge. http://ebookcentral.proquest.com','_blank'

Dikli, S. (2010). The Nature of Automated Essay Scoring FeedbackAuthor(s): Semire Dikli. *CALICO Journal*, *28*(1), 99–134. https://doi.org/10.2307/calicojournal.28.1.99

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

Dörnyei, Z. (2007). *Research methods in applied linguistics : quantitative, qualitative, and mixed methodologies*. Oxford University Press.

Duwairi, R. M. (2006). A framework for the computerized assessment of university student essays. *Computers in Human Behavior*, *22*(3), 381–388. https://doi.org/10.1016/j.chb.2004.09.006

Ellis, R. (1995). Interpretation Tasks for Grammar Teaching. *TESOL Quarterly*, *29*(1), 87–105. https://doi.org/10.2307/3587806

Ellis, R. (1999). *Learning a Second Language through Interaction*. John Benjamins B.V.

Evans, N. W., James Hartshorn, K., & Strong-Krause, D. (2011). The efficacy of dynamic written corrective feedback for university-matriculated ESL learners. *System*, *39*(2), 229–239. https://doi.org/10.1016/j.system.2011.04.012

Fan, N. (2023). Exploring the Effects of Automated Written Corrective Feedback on EFL Students' Writing Quality: A Mixed-Methods Study. *SAGE Open*, *13*(2). https://doi.org/10.1177/21582440231181296

Ferris, D. R. (2003). *Response To Student Writing*. Routledge. https://doi.org/10.4324/9781410607201

Fox, J., & Weisberg, S. (2018). *An R companion to applied regression*. Sage publications.

Gass, S. (2003). Input and Interaction. In Catherine. Doughty & M. Long (Eds.), *Handbook of Second Language Acquisition* (pp. 224–255). Blackwell.

Gass, S. M. (1997). *Input, Interaction, and the Second Language Learner*. Lawrence Erlbaum.

Gass, S., & Mackey, A. (2015). Input, Interaction, and Output in Second Language Acquisition. In B. VanPatten & J. Williams (Eds.), *Theories in Second Language Acquisition: An Introduction* (pp. 180–206). Routledge.

Grimes, D., & Warschauer, M. (2010). Utility in a Fallible Tool: A Multi-Site Case Study of Automated Writing Evaluation. In *The Journal of Technology, Learning, and Assessment* (Vol. 8, Issue 6). www.jtla.org

Hattie, J., & Timperley, H. (2007). The power of feedback. In *Review of Educational Research* (Vol. 77, Issue 1, pp. 81–112). SAGE Publications Inc. https://doi.org/10.3102/003465430298487

Hung, R.-T., Chang, K.-H., Chen, K. T.-C., & Chuang, Y.-S. (2024). *Impact of Automated Writing Evaluation System and Revision Processes on College Students' Writing Skills in English as a Foreign Language Course*. 52. https://doi.org/10.3390/engproc2024074052

Hutchison, D. (2007). An evaluation of computerised essay marking for national curriculum assessment in the UK for 11-year-olds. *British Journal of Educational Technology*, *38*(6), 977–989. https://doi.org/10.1111/j.1467-8535.2006.00686.x

Hyland, K. (2016). Methods and methodologies in second language writing research. *System*, *59*, 116-125.

Hyland, K. (2019). *Second Language Writing*. Cambridge University Press.

Hyland, K., & Hyland, F. (2006a). Contexts and issues in Feedback on L2 writing. In K. Hyland & F. Hyland (Eds.), *Feedback in second language writing: context and issues.* (pp. 1–22). Cambridge university Press.

Hyland, K., & Hyland, F. (2006b). Feedback on second language students' writing. *Language Teaching*, *39*(2), 83–101. https://doi.org/10.1017/S0261444806003399

Karatay, Y., & Karatay, L. (2024). Automated writing evaluation use in second language classrooms: A research synthesis. In *System* (Vol. 123). Elsevier Ltd. https://doi.org/10.1016/j.system.2024.103332

Krashen, S. (1982). *Principles and practice in second language acquisition*. Pergamon Press.

Krashen, S. (1985). *The input hypothesis: Issues and implications*. Longman.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: tests in linear mixed effects models. *Journal of statistical software*, *82*, 1-26.

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

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

Lantolf, J. (2000a). Introducing sociocultural theory. In J. Lantolf (Ed.), *Sociocultural theory and second language learning*. Oxford University Press.

Lantolf, J. (2000b). Second language learning as a mediated process. *Language Teaching*, *33*(2), 79–96. https://doi.org/10.1017/S0261444800015329

Lantolf, J. P. (1999). Second culture acquisition: Cognitive considerations. In E. Hinkel (Ed.), Culture in second language teaching and learning (pp. 28–46). Cambridge University Press.

Lantolf, J. P. (1999). Second culture acquisition. *Second language teaching and learning*, 28-47.

Lantolf, J. P., & Pavlenko, A. (1995). Sociocultural Theory and Second Language Acquisition. *Annual Review of Applied Linguistics*, *15*, 108–124. https://doi.org/10.1017/s0267190500002646

Lantolf, J. P., & Poehner, M. E. (2008). *Sociocultural theory and the teaching of second languages.* . Equinox Publishing Ltd.

Lantolf, J. P., & Thorne, S. L. (2006). *Sociocultural theory and the genesis of second language development*. Oxford University Press.

Li, Z., Dursun, A., & Hegelheimer, V. (2017). Technology and L2 Writing. In *The Handbook of Technology and Second Language Teaching and Learning* (pp. 77–92). Wiley. https://doi.org/10.1002/9781118914069.ch6

Link, S., Mehrzad, M., & Rahimi, M. (2020). Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement. *Computer Assisted Language Learning*, *35*(4), 605–634. https://doi.org/10.1080/09588221.2020.1743323

Loncar, M., Schams, W., & Liang, J. S. (2023). Multiple technologies, multiple sources: trends and analyses of the literature on technology-mediated feedback for L2 English writing published from 2015-2019. *Computer Assisted Language Learning*, *36*(4), 722–784. https://doi.org/10.1080/09588221.2021.1943452

Long, M. (1996). The Role of the Linguistic Environment in Second Language Acquisition. In W. Ritchie & T. Bhatia (Eds.), *Handbook of Second Language Acquisition.* . Academic Press.

Long, M. H. (1981). INPUT, INTERACTION, AND SECOND-LANGUAGE ACQUISITION. *Annals of the New York Academy of Sciences*, *379*(1), 259–278. https://doi.org/10.1111/j.1749-6632.1981.tb42014.x

Mackey, A. (2012). *Input, interaction, and corrective feedback in L2 learning*. Oxford University Press.

Mackey, A., & Gass, S. M. (2015). *Second language research : methodology and design* (Second edition.). Routledge.

McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., & Dai, J. (2015). A hierarchical classification approach to automated essay scoring. *Assessing Writing*, *23*, 35–59. https://doi.org/10.1016/j.asw.2014.09.002

Nunes, A., Cordeiro, C., Limpo, T., & Castro, S. L. (2022). Effectiveness of automated writing evaluation systems in school settings: A systematic review of studies from 2000 to 2020. In *Journal of Computer Assisted Learning* (Vol. 38, Issue 2, pp. 599–620). John Wiley and Sons Inc. https://doi.org/10.1111/jcal.12635

Ohta, A. S. (2000). Rethinking interaction in SLA: Developmentally appropriate assistance in the zone of proximal development and the acquisition of L2 grammar. In J. Lantolf (Ed.), *Sociocultural theory and second language learning*. Oxford University Press.

Ohta, A. S. (2001). *Second language acquisition processes in the classroom : learning Japanese.* . Lawrence Erlbaum.

Page, E. (2003). Project Essay Grade. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective*. Lawrence Erlbaum Associates.

Palermo, C., & Thomson, M. M. (2018). Teacher implementation of Self-Regulated Strategy Development with an automated writing evaluation system: Effects on the argumentative writing performance of middle school students. *Contemporary Educational Psychology*, *54*, 255–270. https://doi.org/10.1016/j.cedpsych.2018.07.002

Pica, T. (1996). Second Language Learning through Interaction: Multiple Perspectives. *Working Papers in Educational Linguistics*, *12*(1), 1–22.

Poehner, M. E., & Lantolf, J. P. (2013). Dynamic assessment. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistic*. Wiley-Blackwell.

Polio, C., Fleck, C., & leder, N. (1998). “If I only had more time:” ESL learners’ changes in linguistic accuracy on essay revisions. *Journal of Second Language Writing*, *7*(1), 43–68. https://doi.org/10.1016/S1060-3743(98)90005-4

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

Powers, D. E., Escoffery, D. S., & Duchnowski, M. P. (2015). Validating Automated Essay Scoring: A (Modest) Refinement of the “Gold Standard.” *Applied Measurement in Education*, *28*(2), 130–142. https://doi.org/10.1080/08957347.2014.1002920

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

Révész, A., Kourtali, N. E., & Mazgutova, D. (2017). Effects of Task Complexity on L2 Writing Behaviors and Linguistic Complexity. *Language Learning*, *67*(1), 208–241. https://doi.org/10.1111/lang.12205

Robinson, P. (2001). Task Complexity, Task Di•culty, and Task Production: Exploring Interactions in a Componential Framework. *Applied Linguistics*, *1*(22), 27–57.

Robinson, P. (2007). Task complexity, theory of mind, and intentional reasoning: Effects on L2 speech production, interaction, uptake and perceptions of task difficulty. *IRAL - International Review of Applied Linguistics in Language Teaching*, *45*(3), 193–213. https://doi.org/10.1515/iral.2007.009

Robinson, P. (2012). Task complexity, cognitive resources, and syllabus design: a triadic framework for examining task influences on SLA. In *Cognition and Second Language*

Instruction (pp. 287–318). Cambridge University Press. https://doi.org/10.1017/cbo9781139524780.012

Roscoe, R. D., Allen, L. K., Weston, J. L., Crossley, S. A., & McNamara, D. S. (2014). The writing pal intelligent tutoring system: Usability testing and development. *Computers and Composition*, *34*, 39–59. https://doi.org/10.1016/j.compcom.2014.09.002

Rudner, L. M. (1992). Reducing Errors Due to the Use of Judges. *ERIC/TM Digest. American Institutes for Research.* www.eric.ed.gov

Rudner, L., & Gagne, P. (2001). An Overview of Three Approaches to Scoring Written Essays by Computer. *Practical Assessment, Research, and Evaluation*, *26*(7). www.eric.ed.gov

Sari, E., & Han, T. (2024). The impact of automated writing evaluation on English as a foreign language learners' writing self-efficacy, self-regulation, anxiety, and performance. *Journal of Computer Assisted Learning*. https://doi.org/10.1111/jcal.13004

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics*, *11*(2), 129–158. https://doi.org/10.1093/applin/11.2.129

Schmidt, R. Schmidt, R.(2001)." Attention." In P. Robinson (Ed.), Cognition and second language instruction (pp. 3-32). Cambridge University Press.

Sheen, Y. (2011). *Corrective feedback, individual differences and second language learning*. Springer.

Shermis, M. D. (2020). International Applications of Automated Scoring. In D. Yan, A. A. Rupp, & P. W. Foltz (Eds.), *Handbook of Automated Scoring* (1st ed.). Chapman and Hall/CRC.

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* (1st ed.). Routledge.

Shermis, M. D., Burstein, J., Higgins, D., & Zechner, K. (n.d.). *Automated Essay Scoring 1 Automated Essay Scoring: Writing Assessment and Instruction*.

Shermis, M. D., Shneyderman, A., & Attali, Y. (2008). How important is content in the ratings of essay assessments? *Assessment in Education: Principles, Policy and Practice*, *15*(1), 91–105. https://doi.org/10.1080/09695940701876219

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford University Press.

Skehan, P., & Foster, P. (2012). Cognition and tasks. In *Cognition and Second Language Instruction* (pp. 183–205). Cambridge University Press. https://doi.org/10.1017/cbo9781139524780.009

Skehan, Peter. (2003). *A cognitive approach to language learning*. Oxford University Press.

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

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. Gass & C. Madden (Eds.), *Input in second language acquisition* (pp. 235–253). Newbury House.

Swain, M. (1995). Three functions of output in second language learning. In G. Cook & B. Seidlhofer (Eds.), *Principle and practice in applied linguistics* (pp. 125–144).

Swain, M., & Watanabe, Y. (2013). Languaging: Collaborative dialogue as a source of second language learning. *The encyclopedia of applied linguistics*, 3218-3225.

Tagliamonte, S. (2012). *Variationist sociolinguistics : change, observation, interpretation*. Wiley-Blackwell.

Tavakoli, P. (2009). Assessing L2 task performance: Understanding effects of task design. *System (Linköping)*, *37*(3), 482–495. https://doi.org/10.1016/j.system.2009.02.013

Truscott, J. (2007). The effect of error correction on learners' ability to write accurately. *Journal of Second Language Writing*, *16*(4), 255–272. https://doi.org/10.1016/j.jslw.2007.06.003

Truscott, J., & Hsu, A. Y. ping. (2008). Error correction, revision, and learning. *Journal of Second Language Writing*, *17*(4), 292–305. https://doi.org/10.1016/j.jslw.2008.05.003

Tuzi, F. (2004). The impact of e-feedback on the revisions of L2 writers in an academic writing course. *Computers and Composition*, *21*(2), 217–235. https://doi.org/10.1016/j.compcom.2004.02.003

Van Beuningen, C. (2010). Corrective Feedback in L2 Writing: Theoretical Perspectives, Empirical Insights, and Future Directions. *IJES*, *10*(2), 1–27. www.um.es/ijes

Van Der Kleij, F. M., Eggen, T. J. H. M., Timmers, C. F., & Veldkamp, B. P. (2012). Effects of feedback in a computer-based assessment for learning. *Computers and Education*, *58*(1), 263–272. https://doi.org/10.1016/j.compedu.2011.07.020

Vygotsky, L. S. (1987). Thinking and speech. *The collected works of LS Vygotsky*, *1*, 39-285.

Vygotsky, L. S., & Cole, M. (1978). *Mind in society: Development of higher psychological processes.* Harvard university press.

Ware, P. (2011). Computer-Generated Feedback on Student Writing. *Source: TESOL Quarterly*, *45*(4), 769–774. https://doi.org/10.5054/tq.201

Warschauer, M. (1997). Computer-Mediated Collaborative Learning: Theory and Practice. *The Modern Language Journal*, *81*(4), 470–481. https://doi.org/10.1111/j.1540-4781.1997.tb05514.x

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. (2013). English language learners and automated scoring of essays: Critical considerations. *Assessing Writing*, *18*(1), 85–99. https://doi.org/10.1016/j.asw.2012.10.006

Wickham, H. (2009). *Ggplot2 elegant graphics for data analysis*. Springer.

Wilson, J. (2017). Associated effects of automated essay evaluation software on growth in writing quality for students with and without disabilities. *Reading and Writing*, *30*(4), 691–718. https://doi.org/10.1007/s11145-016-9695-z

Wilson, J., & Czik, A. (2016). Automated essay evaluation software in English Language Arts classrooms: Effects on teacher feedback, student motivation, and writing quality. *Computers and Education*, *100*, 94–109. https://doi.org/10.1016/j.compedu.2016.05.004

Yu, S. (2021). Giving genre-based peer feedback in academic writing: sources of knowledge and skills, difficulties and challenges. *Assessment and Evaluation in Higher Education*, *46*(1), 36–53. https://doi.org/10.1080/02602938.2020.1742872

Zhang, Zhe Victor, and Ken Hyland. "Student engagement with teacher and automated feedback on L2 writing." *Assessing Writing* 36 (2018): 90-102.

Zhu, M., Liu, O. L., & Lee, H. S. (2020). The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Computers and Education*, *143*. https://doi.org/10.1016/j.compedu.2019.103668