HYBRID FEEDBACK: THE EFFICACY OF COMBINING AUTOMATED AND TEACHER FEEDBACK FOR SECOND LANGUAGE ACADEMIC WRITING DEVELOPMENT

JOHANATHAN WOODWORTH

A DISSERTATION SUBMITTED TO
THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN EDUCATION
YORK UNIVERSITY
TORONTO, ONTARIO

April 2022

Abstract

While researchers and practitioners agree that students need more writing practice and feedback, overburdened teachers often do not have sufficient time to read, mark, and give feedback on students' multiple drafts. Automatic Writing Evaluation (AWE) systems have emerged as a possible solution to give immediate feedback to writers. However, AWE systems lack individualized feedback and feedback on content and can diminish the social and communicative dimensions of writing. Thus, some researchers have advocated that AWE should be construed primarily as a complement to, rather than a replacement of, teacher feedback.

Currently, there is a lack of research on the effectiveness of hybrid feedback, or the combination of teacher and AWE feedback, in the academic writing classroom for supporting the development of second language writing. The current study has started to address this gap by examining if hybrid feedback resulted in differences in approaches to writing, language, content, and organization of writing between a class that received hybrid feedback and a class that received only teacher feedback. A mixed methods design first collected quantitative data and then augmented the quantitative results with in-depth qualitative data. First, pre, post and delayed post-treatment writing tasks were administered to both groups to compare writing in terms of scores and various fine-grained writing indices. A questionnaire on changes in cognitive processes was conducted for both groups, and questionnaire data on the perception of AWE was collected from the experimental group. Second, a focus group interview was conducted as a follow-up to the quantitative stage from the experimental group. A mixed MANOVA comparing changes between and within groups was used to analyze the questionnaire data and changes in writing, and thematic analysis was used to interpret the qualitative data.

The findings suggest that although AWE feedback has limitations, including insensitivity to context, learner needs, meaning, and inability to provide dialogic feedback, combining it with teacher feedback may address some of its limitations, help motivate students to revise and write more often, facilitate autonomous learning, and reduce teachers' workload.

Dedication

I dedicate my dissertation work to Sangjin Han. Without tremendous support from home, returning to school after 15 years would not have been possible. Without warm encouragement and support from Sangjin, this would not have been possible. Thank you for your encouragement, patience, and understanding.

Acknowledgements

Writing this thesis has been a long learning journey for which I owe several people a debt of gratitude. I would like to thank my supervisor Khaled Barkaoui for his continuous support and advice from the inception of the idea to the final dissertation. Much appreciation is also given to the other members of my dissertation committee: Kurt Thumlert and Antonella Valeo.

My appreciation is also sincerely given to the raters Danny Tan and Julia Bae, who worked tirelessly to mark the essays. They helped to clarify the rating process and with whom I bounced the ideas from the inception of the research to the very end of the dissertation. Danny, especially, helped me make sense of the writing when I lost focus and helped me enormously with the mammoth task of proofreading this thesis, thus improving it.

Another thank you goes out to Jenifer Foley from ETS, who was always prompt and helpful with any queries I had for Criterion. The research would have also not been possible without the students who participated in the research and YUELI for allowing the research to take place in their program.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Writing and learning to write require learners to regulate and coordinate multiple component skills across lexical, sub-lexical, syntactic and discourse levels of language (Wilson, 2017). Mastering these skills needs substantial practice and feedback on performance (Storch, 2018); however, this may not be possible in the current climate of English as a Second Language (ESL) classrooms that focus on writing as a product rather than as a process. While multimodal and digital literacy may be the new frontier for second language (L2) instruction, many L2 classrooms focus on traditional written feedback due to institutional and practical constraints (Lotherington & Jenson, 2011). This emphasis, in turn, ignores opportunities for formative assessment and iterative reflection on dynamic writing processes, where risk-taking, experimentation with new vocabulary and structures, and exploring discourse genres and styles are often relegated as being not important and ignores the potential of the new modalities and the resources available.

Classroom writing assessments in ESL classes are dominated by a traditional summative orientation (Lee, 2017). This domination of summative orientation may be due to negative washback, intentional or unintentional influences of testing on the attitudes, behaviour, and the motivation of teachers, learners, and parents (Pearson, 1988), or dominant English proficiency exams, such as the International English Language Testing System (IELTS) and the Test of English as a Foreign Language (TOEFL) (Choi, 2008). Yet, emphasis on writing exclusively as a product for summative evaluation often leads teachers to pay little attention to the writing process, and most assessments are done only summatively (Lee, 2017).

While there is widespread agreement among researchers and practitioners that learners need more writing practice, overburdened teachers have insufficient time to read, mark, and give feedback on learners' multiple drafts and revisions, particularly with large classes (Warschauer & Grimes, 2008). To facilitate giving more feedback, the development of language processing technologies has enabled Automatic Writing Evaluation (AWE) (Burstein et al., 2003). AWE is a broad term for software that typically combines automatic written corrective feedback (AWCF) to promote noticing language errors, with an automated essay scoring (AES) algorithm that evaluates writing quality, a management system to provide learners with multiple drafting opportunities, and a collection of writing resources such as a dictionary, thesaurus, Writers' Handbook and other resources for self-access (Chen & Cheng, 2008; Stevenson & Phakiti, 2014). Typically, AWE systems have a built-in set of topics organized by grade level and writing genres (e.g., descriptive, expository, narrative, persuasive) and have functionality for teachers to add their own topics, add external resources to create integrated writing tasks, create and manage writing portfolios, and other class management tasks (Ranalli et al., 2017). An immediate benefit of AWE is efficacy where feedback is immediate, thus accelerating the practice-feedback loop that is essential in developing metacognitive skills in writing development (Kellogg et al., 2010).

However, while AWE has gained traction, there is a concern that the scoring engine, AES, currently cannot measure the full writing construct, such as the quality of argumentation and content. Other features of writing, such as organization, are only indirectly measured. For example, a study by Shermis et al. (2008b) found that content plays a relatively minor role in the overall score that AWE assigns to essays, ranging from a contribution of 1% of the variance in scores of persuasive and expository essays to about 6% of the variance in scores of descriptive essays. Furthermore, the AWCF is generic and not contingent on learners" needs. It may also devalue the social aspect of writing because learners are writing to a machine. Another danger is that the administrations' promotion of giving feedback via technology may be a step toward making teachers redundant in the classroom. The integration of computer-based instruction may mean the reduction of

teachers' direct labour, which can undermine their autonomy and independence, and decrease their sense of professional identity (Holmes et al., 2021).

Due to these reasons, some researchers like Warschauer and Grimes (2008) have advocated that AWE should complement teacher feedback instead of replacing it. On this account, this study explores a hybrid approach, which combines AWE and teacher feedback that aims to address some of the problems of automated feedback. The hybrid approach may alleviate the time spent on marking and giving feedback to learners while also diminishing the concerns raised against the use of AWE in the EAP classroom. This research emerges from a concern that the integration of machine feedback may not be complemented by an adequate understanding of its impact on learners' writing processes and development. The review of the literature in the next chapter has revealed a scarcity of empirical evidence and longitudinal studies on students' engagement with automated feedback and its impact on their writing processes. The majority of the studies, so far, have compared automated and human feedback, but ecologically, in the classroom setting, the two would be integrated. Therefore, this study examines a hybrid model for feedback in which AWE provides immediate feedback on the lower-level aspects (e.g., lexis, mechanics, grammar, etc.) of students' writing, while teachers give feedback on the higher-level and contextualized aspects of writing (e.g., organization, argumentation, coherence, ideas, etc.). This approach may help teachers to provide students with descriptive feedback and modelling more frequently to support student learning (Cross & O"Loughlin, 2013). By having teachers focus on the higher-level aspects of writing, the hybrid approach can ensure that the aspects of content, argumentation, organization, rhetorical style, etc., are not considered secondary in the L2 classroom. Thus, this study aims to assess the impact of integrating AWE and teacher feedback on students developing useful metacognitive skills for L2 writing, which may help them become independent learners (Wang & Goodman, 2012). Engaging with both types of feedback can promote the cyclical process of review, reflection, and revision which in turn may facilitate students develop' metacognitive strategies to notice, evaluate and improve their academic writing (Zhang & Hyland, 2018).

1.1 Research Questions

Previous research has shown that there is a direct link between feedback and improvements in learners' writing practices (Ferris, John S. Hedgcock, John S., 2013; Myles, 2002). Therefore, this study aimed to examine the changes in students' writing practices after receiving hybrid feedback by comparing the writing practices of students who received a combination of teacher and machine feedback to students who received teacher feedback only before and after receiving feedback using a pretest posttest design. Furthermore, because students are active agents and their perception of feedback is directly correlated with their use and engagement with the feedback (Hyland, 2010), students' views of hybrid feedback were investigated.

Specifically, the study aimed to examine whether and how the use of hybrid corrective feedback and evaluation affects post-secondary ESL students' L2 writing practices compared to students who only receive teacher feedback. Writing practices refer to students' writing approaches and the characteristics of their texts (i.e., grammar, vocabulary, organization, coherence, etc.). The writing approach is operationalized as the five cognitive processes of academic writing as defined by Chan et al. (2017), that is, conceptualization, generating ideas, organizing ideas, generating texts, and monitoring and revising. Additionally, in the hybrid feedback system used in the current study, the AWE system focuses on giving feedback on students' language quality in terms of grammatical competence. In contrast, the teacher feedback focused on higher-level concerns such as strategic and discourse competence, which will be discussed in detail in Chapter 3. The study addressed the following three research questions:

Research Questions

1 How does the use of hybrid corrective feedback affect ESL students" approaches to writing compared to students who only receive teacher feedback?

2 How does the use of hybrid corrective feedback affect the language, content, and

organization of ESL students" writing compared to students who only receive teacher feedback?

3 How do students who receive hybrid AWE corrective feedback view such feedback?

I used a quasi-experimental pre- posttest design to address these questions. Specifically, I taught two classes; in one class, I provided students with hybrid feedback, and in the other, I provided them with teacher feedback only. I then compared the two classes in terms of their writing approaches (i.e., processes) and the language, content, and organization of their essays before and after receiving hybrid or teacher only feedback. Additionally, students who receive hybrid AWE corrective feedback were interviewed about their views of hybrid feedback in addition to a questionnaire about their perception of the AWE system selected for the study.

1.2 Dissertation Structure

The current dissertation is comprised of five chapters. This chapter has introduced some of the terms and background of automated feedback, the significance and rationale of the study's focus on a hybrid approach to feedback, and the research questions. Chapter 2 critically reviews the literature on the role of feedback in L2 writing development and automated feedback in the classroom and builds a case for using a hybrid approach to feedback. Chapter 3 provides detailed information on the study's research context, design, participants, instruments, data collection, and data analysis procedures. Chapter 4 presents the results concerning the impact of a hybrid feedback approach on students" writing processes and texts and their perception of hybrid corrective feedback. Chapter 5 summarizes and discusses the results, the limitations, and the theoretical, pedagogical and research implications of the study.

Chapter 2: Literature Review

This chapter examines leading writing theories and models and their relations to feedback research, the role of AWE in the classroom, and previous research on hybrid feedback. It first discusses the role of assessment in the L2 classroom. Then, theoretical perspectives on the role of feedback within theories of writing are discussed to highlight the importance of feedback in L2 writing development. Next, methodological challenges of researching WCF in SLA and L2 writing research are discussed. The following section reviews AWE feedback research and concerns about AWE feedback use and validity. The last section describes a hybrid approach to feedback intended to address validity concerns with using AWE feedback in the classroom.

2.1 The Role of Assessment in the L2 Classroom

Assessment in the L2 classroom can serve different purposes. Traditional testing serves the purpose of assessment of learning (AoL). AoL measures and documents student performance against a specified standard that is typically formal academic discourse and associated norms. In many L2 classrooms, this is the predominant use of assessment. In contrast, assessment for learning (AfL) focuses on enhancing learning by identifying weaknesses and strengths in students" performance to improve teaching and learning. Lastly, assessment as learning (AaL) extends the role of AfL and is "a process through which pupil involvement in assessment can feature as part of learning" (Dann, 2002, p. 153). In both AaL and AFL, feedback plays a more important role than in AoL. While all three uses of assessment attend to different goals in the classroom and are informed by different orientations to learning, ideally, they should work in conjunction to facilitate learning and, with AaL in particular, student metacognition and increased autonomy as students are more directly involved in responding to, and learning from, feedback from instructors, peers, comparative models, technology tools, or other means.

In the literature, AoL is often used interchangeably with summative assessments that

serve administrative and certification purposes because it usually occurs at the end of a specific learning period. AoL"s role is fundamentally normative to compare student performance to a set standard or to other students to make decisions about student progress or achievement. In the L2 classroom, Lee (2017) suggests that in traditional AoL practice, "teachers simply assign the topic without providing specific learning targets; student writing is assessed against a general assessment Criterion such as content, language, and organization; teacher feedback is summative rather than formative, mainly comprising of corrections in language forms" (p. 14). In many L2 classrooms, the ultimate goal of the class is the product - the student-produced texts. Feedback is only given on the last draft, which obviates the need for students to use teacher feedback for reflection, problem-solving, looking at provided annotations and instructor exemplars, and improvement. A study of students" perception of summative assessment by Maclellan (2001) showed that students do not take advantage of assessment to improve learning because summative assessments do not foster the skills of monitoring and regulating the quality of students" learning and students may not reflect on or care about feedback because they perceive that learning is "over" (Lee, 2017).

Unlike AoL, which usually occurs at the end of the learning process, AfL shifts from making judgments about students" performance to diagnosing their strengths and weakness to enable the monitoring of learning and progress, thus, facilitating the improvement of both learning and teaching. AfL is often referred to as formative assessment. Finally, AaL, which is a subset of AfL, emphasizes student-centred reflection, inquiry, and learning. AaL encourages students to monitor and reflect on their learning processes and progress and make adjustments to their writing and learning accordingly. While AoL serves a purpose, scholars of L2 writing like Lee (2017) promote AfL and its subset, AaL, because they believe that "classroom writing assessment should reflect a real, substantive focus on the improvement of learning and teaching" (p.12). Further, where AoL often summarily "punishes" error and risk-taking, AfL/AaL both enable students to respond to feedback and learn from errors, which is a significant aspect of revision, self-directed learning, and metacognition.

In contrast to AoL, both AfL and AaL draw on a social constructivist framework that combines cognitive and sociocultural theories (Shepard, 2000). Metacognition, thinking about thinking, or cognitive theory plays a vital role in AfL and AaL. Being able to solve problems within each domain of practice involves what Sternberg (1992) termed "executive processes" such as (a) recognizing the existence of a problem, (b) deciding on the nature of the problem, (c) selecting a set of lower-order processes to solve the problem, (d) developing a strategy to combine these components, (e) selecting a mental representation of the problem, (f) allocating one"s mental resources, (g) monitoring one"s problem solving as it is happening, and (h) evaluating problem-solving after it is done (Shepard, 2000, p. 21). Second, adept students are able to take charge of their own learning using a variety of self-monitoring processes (Brown, 1994). This process requires (a) goal setting, (b) self-monitoring with reference to the goal, (c) interpreting and utilizing feedback (e.g., from teacher and peers) that results from self-monitoring, and (d) modification of goal-directed action (e.g., adjusting or redefining the goal) (Andrade et al., 2012). Lastly, the theory of motivation shows that students are more likely to be motivated by intrinsic rather than extrinsic goals (Deci & Ryan, 1985). That is, students are more likely to be motivated by the intrinsic desire to attain mastery or further a skilled practice, and achieve competence rather than extrinsic performance goals.

While some scholars (e.g., Lam, 2016; Lee, 2011; Lee, 2017) believe that AaL is a subset of AfL, others see enough distinctions to see two separate approaches (Sadeghi & Rahmati, 2017). For example, AaL draws upon theories of motivation, autonomy, metacognition, and self-regulation. The student is active in self-assessment where they judge the quality of their work and amend/revise as necessary, responding to the feedback of peers, instructors, or external compositional models. AaL stands in contrast to AfL where the reflection of learning is provided by "teachers rather than being obtained in a process of meta-cognitive engagement" (Sadeghi & Rahmati, 2017, p. 51).

While traditional standardized assessments of learning tend to view writing as a product only, theory and research suggest that the focus should be on the process of writing. This means that L2 learners should be actively engaged in learning and reflection by engaging in self-assessment, leveraging their intrinsic motivation, and setting their own goals to achieve proficiencies in writing (Lee, 2017). In this approach, teachers should act as facilitators and "just-in-time" resources to help students monitor their own learning and writing by providing continuous feedback and taking advantage of the complementary functions of AoL, AfL, and AaL.

Although many ESL classrooms view writing as a product to help students develop the cognitive and practical skills of writing and to foster language learning and acquisition, current research suggests that the focus should be on the process of writing. AaL and AfL support this approach by drawing on motivation, autonomy, metacognition, and self-regulation theories. WCF is seen as an essential part of this process because it mediates language acquisition by serving a metalinguistic function of reviewing so that the students can reflect on their use of the target language and enhance their awareness of forms and rules in the process of turning declarative knowledge into procedural knowledge.

2.2 Writing Feedback in the ESL Classroom

Research on feedback initially stemmed from L1 research on writing. Much of the original L1 writing research examined the writing process and found that good writers revised more often than poor writers (Stallard, 1974), and that good writers revised content more often than form (Faigley & Witte, 1981; Perl, 1979; Sommers, 1980). These findings suggested that teachers should focus on content more than form (Fathman, 1990). These findings, however, may not apply to L2 writing. For example, regardless of how well-rounded and organized the thoughts are, L2 writing samples typically contain sentence and discourse-level errors, impeding communication. Studies have shown that a lack of grammatical accuracy in ESL student writing may hinder academic success in

universities (e.g., Evans & Green, 2007; Evans & Morrison, 2010; Ferris, 2002). Consequently, the majority of research indicates that attention should be paid to both the content and form of L2 learners" writing (Ferris, 2002).

However, the use of WCF in second language writing classes has been the subject of much controversy. Proponents believe that WCF is an integral part of writing development supported by various theoretical perspectives on writing (Ferris & Roberts, 2001; Ferris, 1999; Ferris, 2003). However, others (e.g., Truscott, 1996; Truscott, 2007) believe that grammar correction is ineffective in facilitating improvement with respect to writing and that time spent on WCF could be better used for classroom instruction instead.

The question of the effectiveness of feedback on L2 writing is further complicated by the fact that research on L2 feedback has been conducted from two different perspectives, Second Language Acquisition (SLA) and L2 writing, that differ in terms of the questions they ask, the types of feedback they focus on, and the research designs they use. Although key theories informing SLA and L2 writing researchers are often the same, the design of their research is different. While L2 writing and SLA researchers often examine similar phenomena, they often do not necessarily ask the same research questions. SLA-focused researchers investigate whether WCF facilitates the acquisition of particular linguistic features (e.g., definite articles). In contrast, "L2 writing researchers generally emphasize the question of whether written CF helps student writers improve the overall effectiveness of their texts" (Ferris, 2010, p. 181). Thus, SLA-oriented studies have aimed to find out whether WCF facilitates the long-term acquisition of particular linguistic features, including rule-governed features such as article usage, verb tense, and subject-verb agreements and non-rule governed features such as prepositions, collocations, and word choice (Wang & Jiang, 2015). L2 writing studies usually do not limit the type of WCF given. In other words, SLA researchers generally focus on feedback on a few carefully selected error types, while L2 researchers generally look at comprehensive feedback, including feedback on language, content, coherence, etc.

In addition, the research designs of the two traditions are also very different. A meta review by Ferris (2010) found that L2 writing studies are usually set within writing classrooms; there may or may not be a control group or a pretest-posttest design; some studies do not define or delimit which types of student writing errors received written CF; and there is variation as to how written CF is provided. In contrast, the SLA studies are conducted under far more controlled experimental conditions by employing comparison and experimental groups and use pretest, posttest, delayed test designs (Ferris, 2010). While some researchers have criticized L2 writing research for its less controlled and inconsistent designs (Bitchener, 2008; Sheen, 2007; Truscott, 1996; Truscott, 2007), others question if "in the interest of empirical rigour, some of the SLA research efforts on written CF have been so narrowly focused that it would be difficult to transfer their approach and findings to a real writing classroom or to a diverse group of students" (Ferris, 2010, p. 186). In other words, L2 writing researchers believe that helping students improve linguistic control involves a variety of approaches, and they would not likely focus on the acquisition of a specific linguistic feature.

Furthermore, research on feedback can be guided by different theoretical frameworks. The following paragraphs will briefly overview the cognitive perspective to feedback and L2 acquisition, a traditionally popular theoretical case supporting WCF, and the sociocultural perspective, a more recently adopted framework to examine WCF.

2.2.1 Theoretical Perspectives on the Role of WCF

Theoretical accounts of language acquisition via cognitive processing have sought to explain the nature of L2 knowledge and the cognitive processes involved in its development (Bitchener & Storch, 2016). In the beginning of the 1980s, Krashen (1984) produced what has come to be regarded as the first comprehensive theory and model of SLA (Bitchener & Ferris, 2012). Although this theory has received considerable criticism over the years, it was nevertheless influential in shaping the research direction for the last

30 years. Krashen makes a distinction between 'acquisition' and 'learning. He posits that they are separate processes: equating 'acquisition' with implicit knowledge and 'learning' with explicit knowledge. Because these processes are different and cannot become integrated, he saw no value for acquisition that results from WCF. He believed that when learners are exposed to sufficiently rich comprehensive input, knowledge of language is unconsciously acquired, negating the role of WCF in L2 instruction. However, DeKeyser (2007) disagreed, arguing instead that there is an interface between the two processes: 'learning' occurs when instruction focuses on form, which could be provided through WCF, and learned knowledge could be converted into 'acquisition' when learners interact in meaningful communication.

Another influential hypothesis, the Interaction Hypothesis, stems from observations of speaking interaction. Long (1996) claimed that conversational moves such as recasts and clarification requests provide learners with a primary source of language input, enabling them to negotiate meaning in a natural flow of conversation, thus facilitating language development. Long (1996) explained that negotiation triggers interactional adjustments, which facilitate acquisition because it connects input, internal learner capacities, and output in productive ways. During a negotiation event, the role of attention is elevated. To put it another way, "attention, accomplished in part through negotiation, is one of the crucial mechanisms in this process" (Gass, 1997, p. 132). Long (1981) proposed that "environmental contributions to acquisition are mediated by selective attention and the learner's developing L2 processing capacity, and that these resources are brought together most usefully, although not exclusively, during negotiation for meaning. Corrective feedback obtained during negotiation work or elsewhere may be facilitative of L2 development, at least for vocabulary, morphology, and language-specific syntax, and essential for learning certain specifiable L1–L2 contrasts" (p. 414). In Long's view, both comprehensible input and L2 development stem from modifications made by L2 learners when communicating, and feedback facilitates this process (Mackey et al., 2012). In other words, confirmation checks, comprehension checks, and clarification requests, which WCF can facilitate, help resolve communication difficulties.

While early versions of the Interaction Hypothesis incorporated Krashen's claims about comprehensible input being necessary, many researchers like 1995; Swain (1985) argued that while comprehensibility is necessary, it is not sufficient. From her research experience in the French immersion context, Swain (1993) found that despite years of exposure to sufficiently rich comprehensible input in communicative classrooms, students lacked grammatical accuracy regarding morphology and syntax (Harley & Swain, 1984; Lightbown & Spada, 1994). Swain (1985) argued that comprehensible input alone is not enough for learners to produce grammatical and error-free utterances. Furthermore, one of the most important reasons for promoting output to improve second language learning is that when learners experience communication difficulties, they need to be pushed into making their output – that is, communicative action and situated doing or making - more precise and appropriate (Shao, 2015, p. 160). Likewise, Swain (1985) has noted that "producing the target language may be the trigger that forces the learner to pay attention to the means of expression needed in order to successfully convey his or her own intent" (p. 249). Beyond "exposure", Swain (1985) suggested that second language learners need to be urged to produce output, arguing that "being pushed in output, it seems to me, is a concept parallel to that of the i + 1 of comprehensible input. Indeed, one might call this the comprehensible output" (p. 249).

Swain (1993; 1995) has refined her hypothesis and termed it the Output Hypothesis and specified the following four functions of output. First, output has a fluency function that provides learners with opportunities for developing speedy access to their existing second language knowledge in the actual use of language in meaningful contexts. Second, output has a hypothesis-testing function. In the process of producing output, learners are able to form and test their hypotheses about the comprehensibility and linguistic accuracy of their utterances in response to feedback obtained from their interlocutors or readers. Third, output has a meta-linguistic function. It is claimed that "as learners reflect upon their own target language use, their output serves a meta-linguistic function, enabling them to control and internalize linguistic knowledge" (Swain, 1995, p. 126). In other words, output processes enable learners to reflect upon their target language use and

consolidate their linguistic knowledge about the grammatical features they already have declarative knowledge of: for example, abstract grammatical rules. Reflection on language through use and practice may enhance their awareness of forms, rules, and form-function mapping in a meaningful context. Finally, output serves a noticing function. Namely, in producing the target language in writing, "learners may notice a gap between what they want to say and what they can say, leading them to recognize what they do not know or know only partially" (Swain, 1995, pp. 125-126). The recognition of problems may prompt the learners to selectively attend to the relevant information in the input, which will trigger their interlanguage development (Shao, 2015, p. 161). WCF could be an important facilitator in helping learners to attend to input selectively, especially as learners are enabled to make connections between what they are specifically doing, saying, or writing, and the feedback, models and new "inputs" they receive in situ. Swain (2000) believes that metalinguistic discussions of the language itself during the writing process may contribute to language learning, and this is precisely what WCF aids.

Taking the information processing approach of the Output Hypothesis as the starting point, De Bot (1996) argues that output serves an essential function in second language acquisition, precisely because it can generate highly specific input that the cognitive system needs in order to build a coherent set of knowledge. He claimed that output plays a direct role in turning declarative knowledge into procedural knowledge from an information processing perspective when learning an additional language. Furthermore, De Bot (1996) argues that when the learner's output does not match the correct form, corrective feedback will allow the learners to pay attention temporarily to language form instead of meaning, which could prevent the solidification (further repetition) of the erroneous form in memory or in embodied habit. In other words, on the one hand, output invites feedback that promotes noticing. On the other hand, feedback plays an indispensable role in inciting learners to produce grammatically more accurate output, which may consolidate already-learned knowledge of the rules, enhance form-meaning mapping, or trigger faster access to the already-learned structure to develop automaticity (Shao, 2015, p. 161).

The Noticing Hypothesis grew out of points made about the role of attention in the Interaction Hypotheses. Schmidt (2001) separates 'noticing' from 'meta-linguistic awareness" by "assuming that the objects of attention and noticing are elements of the surface structure of utterances in the input – instances of language, rather than any abstract rules or principles of which such instances may be exemplars" (p. 5). He sees attention to input to be essential for storing that information in memory and a necessary precursor to hypothesis formation and testing in L2 development. Furthermore, this noticing can originate internally or externally: internally driven noticing is when the input becomes noticeable to the learner because of internal cognitive changes and processes and externally derived noticing occurs "when input becomes more noticeable because the manner of exposure is changed" (Schmidt, 2001, p. 10). In the L2 classroom, WCF may facilitate noticing because when the teacher instructs or gives feedback (corrections, alternative modeling, etc.), external noticing occurs. This helps to focus students' attention on the forms and the meanings in the input that facilitates students" refection on their own output, which, as Schmidt notes, is a prerequisite for subsequent processing.

Schmidt (2001) claims that noticing is a necessary condition for the storage of new forms. In other words, learning and memorization require sustained attention and awareness of correct forms. He maintains that there are specific factors that determine what is noticed: 1) expectations, 2) frequency, 3) perceptual salience, 4) skill level, and 5) task demands (Shao, 2015, p. 162). In this hypothesis, attention is seen as being limited, selective and essential. Because attention is limited, an activity that draws upon it will interfere with the current focus, and attention must be strategically allocated. Thus, attention is subject to voluntary control to pay attention to one stimulus over another. In the L2 classroom, the teacher helps students attend to different aspects of the target language and focus on an important function. Due to these processes, attention is essential for learning. Schmidt (1995) posits that attention is necessary for input to become available for further mental processing.

Empirical evidence supports this claim. Studies (e.g., Jeon & Kaya, 2006; Kasper &

Rose, 2002) have shown that explicit instruction that focuses on noticing forms tends to be more effective than implicit instruction. This may be because aspects of the salient and meaningful input are typically those that draw the learners' attention. For example, features that lack salience or communicative value may not be noticed, and L2 learners may benefit from having their attention drawn to formal features of the target language. While Krashen (1984) stated that it was sufficient for learners to pay attention to input to acquire language, Schmidt (1995) posited that learners need to pay attention to both linguistic form and grammatical structure if the acquisition is to occur. As such, controlled activities like corrective feedback can facilitate the conversion of declarative knowledge into automatized procedural knowledge (Bitchener & Ferris, 2012).

Drawing upon the Output and Noticing Hypotheses, Long (1996) put forward a reformulated version of the interactive hypothesis. He stated that "negotiation for meaning, and especially negotiation that triggers interactional adjustments by the native speaker or more competent interlocutor, facilitates acquisition because it connects input, internal learner capacities, particularly selective attention, and output in productive ways" (Long, 1996, p. 451). The combination of these theories and their "constellation of features - interactionally modified input, having the learner's attention drawn to his/her interlanguage and to the formal features of the L2, opportunities to produce output, and opportunities to receive feedback" (Mackey et al., 2012, p. 9) are the core components of the reformulated Interaction Hypothesis (Long, 1996). The hypothesis is well established, and it has encouraged research investigating not only whether interaction impacts L2 learning but also "(a) which aspects of the L2 benefit the most from interaction; (b) how individual difference variables mediate the relationship between interaction and L2 development; and (c) what forms of interaction (and in particular, what types of feedback) are the most beneficial for L2 learners (how various types of interactional feedback differentially impact various L2 forms)" (Mackey et al., 2012, p. 10). In summary, in the cognitive tradition, explicit declarative L2 knowledge can be proceduralised through meaningful contextualized practice with WCF, which facilitates noticing and is converted to implicit, acquired knowledge (Bitchener & Storch, 2016).

However, the cognitive model discounts the fact that in order for learners to really learn, attention cannot be disconnected from the actual practices, purposes, actions, and interests of actual people and their interactions. Sociocultural theory (SCT) addresses some of these issues.

Unlike the cognitive model, SCT encapsulates learning within human interaction. Ellis (2010a) suggested SCT may be best equipped to explain CF as a sociocognitive phenomenon. Sociocultural theory is based on the work of Vygotsky (1980). The fundamental premise of the theory is that learning occurs during interactions between an expert and a novice mediated by artifacts that may be physical (material artefacts or electronic devices) or symbolic (language, writing, multimodal semiotic artefacts). Within SCT, three interrelated constructs are the most relevant to discussions of WCF: Zone of Proximal Development (ZPD), which explains the effectiveness of WCF, mediation tools, which frame the discussion on how feedback is delivered and processed, and activity theory, which proposes that any activity such as learners' uptake of WCF needs to be considered holistically (Bitchener & Storch, 2016).

ZPD is defined as the distance between what a learner can accomplish alone and what that learner can achieve with the support of more capable experts, peers, and/or cultural artifacts (Vygotsky, 1987). Therefore, L2 acquisition informed by SCT focuses on the nature of assistance that a person with more proficient skills offers to a person with less proficient skills or how the learner utilizes artefacts to mediate the differences in skills. However, not all aid is appropriate. For example, too much help can inhibit development, and too little can lead to frustration. Several approaches to provide the most effective amount of assistance are offered, such as one by Feuerstein and his colleagues (Feuerstein et al., 1979; Feuerstein et al., 1980; Feuerstein et al., 1988; Feuerstein, 1990) who have developed the concept of Mediated Learning Experience (MLE). Feuerstein et al. (1988) have provided three criteria for a mediated learning interaction: intentionality/ reciprocity, transcendence, and meaning. While the original intent of the criteria was to shape student/teacher interaction, it can be applied to AWE. Intentionality refers to the

deliberate effort of receiving and giving feedback, and reciprocity refers to interactions where students are actively involved in the feedback process rather than being passive recipients. Transcendence refers to students' ability to use the learning gained from one feedback process in new contexts. In other words, students can apply their learning across different writing tasks. Finally, meaning conveys the significance of the learning process for the student. In MLE, the goal is to provide the type of support that will enable the student to perform beyond their current level of proficiency using these means or criteria.

The mediation tools (in this study, the combined feedback from AWE and instructor) may impact the nature of the assistance provided and the learners' response to the assistance (Bitchener & Storch, 2016). In SCT, mediation tools are separated into symbolic or physical. Of all symbolic tools, language itself is considered a primary mediation tool because it mediates the interaction between people. However, digital mediation tools have become increasingly important with the rapid developments of education technology. There have been a number of new means of delivering feedback, such as computer-facilitated feedback that may be synchronous, delayed, or asynchronous. However, these various tools for mediation have limitations or affordances in different contexts that may contribute to engagement with feedback. Thus, examining the effectiveness of feedback needs to consider how it is delivered and how learners engage with it.

Activity Theory aims to understand the capabilities of a single individual through analyzing the cultural and physical aspects of human actions (Bertelsen & Bødker, 2003). Although there have been many revisions of activity theory, they all focus on the specific activity rather than on any individual. The underlying premise of the theory is that to understand a situation or activity, the behaviour of all individuals, including artefacts and interactive systems involved, and the role of mediation tools that facilitate the activity need to be examined (Bitchener & Storch, 2016). In other words, the theory can be a material basis to analyze how people in socially organized systems such as schools acquire "complex abilities such as writing and languages" (Cumming, 2015, p. 77). In

summary, SCT justifies the use of WCF for L2 development. In the L2 classroom, one activity is teachers negotiating the ZPD together with their students to extend L2 writing tasks with increased mutual understanding, independence, and effectiveness by scaffolding and giving feedback. Students can negotiate as well with tools, peers, and artefacts and experiences in and outside of the classrooms. Thus, sociocultural theory sees learning as "socially and culturally constructed, with learners shouldering the responsibility of learning and the teacher playing the roles of a facilitator" (Lee, 2017, p. 12). In other words, WCF is effective when it is tailored to the learners' needs. Through scaffolding, learners can learn to use the target language with assistance from teachers or peers and/or tools in the classroom to produce language that they would not yet be able to produce independently (Sheen, 2010). The mediation tools, both symbolic and physical, must be considered in how they facilitate engagement and the activities (learning challenges) of language acquisition; the individuals who take part in the activity and the mediation tools need to be considered simultaneously.

The exploration of how feedback affects L2 learning, acquisition, and development is central to feedback research. Corrective feedback in L2 learning has been of considerable interest to SLA researchers since 1995; Swain (1985) contended that comprehensible input is a necessary but not a sufficient condition for learners' L2 development. In the process of L2 acquisition, input, interaction, monitoring, and noticing may be instrumental in the interactionist perspective of learning as suggested by sociocultural theory. The adoption of sociocultural theoretical perspectives on L2 writing acquisition has encouraged research that examines how learners use language in interaction during learning activities (Wigglesworth & Storch, 2012) and sharply contrasts with the cognitive-interactionist perspective (Bitchener & Storch, 2016). Sociocultural theories maintain that learning is essentially a social process rather than one limited to within the individual, too often seen or theorized as a kind of cognitive "processor" of inputs and outputs, abstracted from history, culture and diversity. Learning, including L2 learning, develops in the social, inter-mental plane, and only subsequently it is appropriated by the individual into the 'intramental' plane (Vygotsky, 1987). Feedback on students' writing,

for instance, involves corrective interaction between students, teachers and/or peers, and/or tools and artefacts, and such interaction occurs in order to negotiate meaning within a social context, as well as negotiate increasingly complex learning challenges; the external input such as WCF is meaningful to students (Brown, 2000). In other words, mediation of WCF provided by teachers can promote positive change in students" use of linguistic resources. Moreover, WCF by teachers is dialogic because it is ongoingly and interactively adjusted for the ZPD of students, insofar as a teacher or a tool can assess those various "zones" and provide further scaffolding for learning without constraining the student or impeding self-directed problem solving.

2.2.2 Previous Research on Written Corrective Feedback (WCF)

A review of WCF research conducted within the cognitive and SCT perspective suggests that WCF facilitates L2 development. In the cognitive perspective, while classroom practice is predicated on the assumption that feedback can improve learning (Hyland & Hyland, 2006), the literature is conflicted on the effectiveness of different types, immediacy, and amount of feedback in its findings. Ellis (2009) writes that "there is no widely accepted theory of grammatical complexity to help teachers (or researchers) decide which rules are simple and portable or to determine which features are marked" (p. 6). In addition, while written corrective feedback (WCF) is seen as a central aspect of an ESL writing program, "the research literature has not been unequivocally positive about its role in the classroom" (Hyland & Hyland, 2006, p. 1).

As a result of the call to reject written grammar correction by Truscott (1996), a growing body of research has investigated whether WCF can facilitate improved accuracy, what conditions need to be met for this to occur, and which approaches to WCF are more effective (Bitchener & Knoch, 2015, p. 407). Accordingly, later research has disputed the abandonment of the feedback process. For example, longitudinal research measuring improvement in writing after error feedback shows that students improved accuracy (Bitchener & Knoch, 2009; Ferris & Helt, 2000; Ferris & Roberts, 2001; Sheen, 2007; Van Beuningen et al., 2012). However, so far, SLA research has been limited to testing

WCF effectiveness with specific linguistic error domains and categories (Bitchener & Knoch, 2010).

In contrast, L2 writing studies have examined the overall qualities of writing; due to limitations in the design and execution of these studies, it is difficult to assess their claims. For example, only text revisions were measured, without attention to improvements in new texts, and studies did not have pretest measures, nor did they utilize different instruments in pre and posttests (e.g., Bitchener & Knoch, 2008). In addition, many L2 writing studies did not have comparison groups to evaluate WCF improvements in accuracy (Chandler, 2000; Chandler, 2003; Ferris et al., 2000; Ferris, 1995a; Ferris, 1997; Lalande, 1982).

Furthermore, current literature has shown that there are conflicting perceptions about WCF practices. For example, some studies indicate that teachers are overly concerned about grammar (Ferris & Roberts, 2001; Hyland & Hyland, 2006; Lee, 2005; Lee, 2014b; Lee, 2017; Robb et al., 1986; Zamel, 1985), while other studies contradict this (Sheen, 2007; Truscott & Hsu, 2008; Truscott, 1996; Truscott, 2007). Some researchers have proposed that teachers believe the WCF they give is effective (Bitchener, 2008; Kepner, 1991; Zamel, 1985) while others state that teachers are doubtful of such effectiveness (Hyland, 1998; Hyland, 2013). Some studies suggest that teachers are fundamentally unsure if WCF has a positive effect (Guénette, 2007; Hyland & Hyland, 2006; Kepner, 1991; Lee, 2004; Lee, 2005; Truscott & Hsu, 2008; Truscott, 1996; Truscott, 2007), while other studies have implied that teachers are inconsistent and arbitrary with their comments (Cohen & Cavalcanti, 1990; Cohen, 1987; Lee, 2004; Lee, 2005; Nystrom, 1983; Zamel, 1985).

Another issue concerns the design of studies of WCF. Earlier studies on WCF were not conclusive about feedback effects on writing development. Researchers like Ferris (2004) and Guénette (2007) attribute this to poor research design and lack of comparability. Previous studies of WCF can be categorized broadly in two bodies: improvements of

accuracy of a particular text as examined by Ashwell (2000), Fathman (1990), Ferris (1997), and Ferris and Roberts (2001) and improvements of accuracy in new texts (Bitchener, 2012). Researchers have criticized the first body of work as not necessarily illustrating evidence of learning; "evidence of learning can only be seen when accuracy in one or more new texts is compared with inaccuracy in an earlier text" (Bitchener, 2012, p. 353). Thus, the differences in earlier findings compared to newer, more experimental studies may be due to 1) flaws in the design, gathering and analysis of data, 2) different design variables, and 3) differences in research designs. The later studies, which were conducted under more controlled experimental conditions that focused on a few carefully chosen and defined error types, and feedback provided systematically for both revisions and new pieces of writing, have shown improved accuracy in both immediate and delayed posttests (Bitchener & Knoch, 2008; Bitchener & Knoch, 2009a; Bitchener & Knoch, 2010; Bitchener, 2008; Ellis et al., 2008; Sheen et al., 2009; Sheen, 2007).

The evidence on whether feedback has an influence on student revision seems "conclusive". In a study by Fathman (1990), the authors showed that every student who received grammar feedback received higher grammar scores on their revised drafts. Moreover, empirical research shows that error feedback on drafts helps L2 learners revise their texts. For example, research designs with experimental groups showed that the treatment group(s) outperformed the experimental groups on drafts (Ashwell, 2000; Ferris & Roberts, 2001; Truscott & Hsu, 2008). This is supported by both immediate and delayed posttests supporting the claim that WCF can result in improved accuracy over time (Ferris, 2011).

Much of the previous research on WCF focused on the dichotomy between direct and indirect feedback. Indirect feedback only indicates that an error has occurred. Examples of indirect feedback can be circling or underlining the error with or without correction codes, critical annotations, or alternative modelling (reformulation). In contrast, Direct CF provides learners with the corrected versions of linguistic structures. Researchers who support indirect feedback suggest that it may engage learners more by directing them to

be more reflective and analytical about their errors to fix them by determining the error and the correction. Proponents of direct CF suggest that, especially for less proficient learners, it can reduce confusion and provide them with information to resolve more complex errors that the learners may not resolve independently.

Earlier research on the effects of direct and indirect CF has shown that there were no statistically significant differences between the two types of feedback (Lalande, 1982; Robb et al., 1986; Semke, 1984), but later studies seem to have conflicting results. Some research indicated that while direct error correction led to a higher percentage of correct short-term revisions, indirect corrective feedback may be more effective in contributing to the increase of long-term writing development (e.g., Ferris, 2006; Van Beuningen et al., 2012; Bitchener & Knoch, 2010) have shown that while short-term effects were similar, direct error correction had greater long-term gains than indirect CF. Bitchener and Storch (2016) argue that the conflicting findings may be due to design differences and possible variables that may have impacted the findings: such as the impact of direct and indirect WCF for different proficiency groups, mediating individual factors, and the possibility of certain types of errors being more responsive to different types of CF could all be confounding variables. For example, Bitchener et al. (2005) found that WCF was useful for uptake and retention for articles and simple past tense usage but not for prepositions. Moreover, researchers have suggested that the differing findings may be due to the proficiency level of the learners; learners with low language proficiency are less likely to benefit from indirect feedback because they do not have a sufficient level of linguistic competence to be able to self-correct their errors (Bitchener & Knoch, 2009a; Bitchener & Knoch, 2010; Ferris & Roberts, 2001; Ferris, 2004; Ferris, 2011). In contrast, indirect feedback may be more suitable for learners with higher-level proficiency because they have more cognitive resources to decipher the feedback and self-correct (e.g., Ferris, 2006; Ferris, 2010; Lalande, 1982).

In addition, Ellis et al. (2008) have opined that the distinction between direct and indirect WCF is problematic because when teachers give indirect WCF, they assume that the

learners already know the structure of the language and can self-correct in response to the feedback. In other words, "indirect feedback can only lead to an increase in control of a linguistic form that has already been partially internalized. It cannot lead to new learning (i.e., learning of new linguistic forms)" (Storch, 2010, p. 40). Due to the difficulty in determining if a structure is new or needs more practice, Ellis et al. (2008) have suggested that the distinction between direct and indirect feedback is not worth investigating.

Another issue concerns which error categories are amenable for WCF. Studies have shown that some errors are less treatable due to not having systematic or teachable rules (Bitchener, 2008; Ferris, 2006; Xu, 2009). Ferris (1999) has made a distinction between treatable and untreatable errors. For instance, Ferris (2010) lists "word order, sentence boundaries, phrase construction, word choice, or collocations" (p. 193) as untreatable. Van Beuningen (2011) identified treatable errors as rule-governed and untreatable errors as non-rule-governed errors. For example, article usage, verb tense, and subject-verb agreement have clear rules. These errors are more amenable for WCF. Thus, a large number of studies that investigate WCF have examined a narrow set of rule-governed structures like article usage Van Beuningen (2011). However, tightly controlled studies such as these, which are time-restricted, with impromptu tasks that are not integrated with other skills, lack ecological validity because they do not reflect real-life writing scenarios (Polio, 2017). Likewise, Xu (2009) has argued that such a narrow focus does not help to explain how learners" control over other linguistic structures might be affected by the treatment.

Due to the application of SCT in L2 acquisition and WCF research being emergent, there are fewer empirical SCT oriented studies of WCF. There have been two foundational studies on the effectiveness of scaffolded and non-scaffolded CF. Studies by Nassaji and Swain (2000) and Erlam et al. (2013) both provided targeted feedback in oral conferences on writing. Both studies found that scaffolded feedback encouraged learners to self-correct and feedback that is scaffolded and sensitive to learners" ZPD has a greater

impact on learning, but such results may not be generalizable due to intensive one-on-one sessions being impractical in regular class settings. Although, there is a limited body of research on scaffolded feedback on writing development, the literature suggests that feedback attuned to learners" ZPD aids in the acquisition of L2 writing in drafts and new pieces of writing compared to groups that did not receive such contingent feedback.

Studies that examined the use of tools focused on either symbolic tools (language) or materials tools (technology) have also been few. Many studies that examined how language mediates the processing of WCF compared reformulation as feedback to other forms such as joint writing, stimulated recall, and revisions (Swain & Lapkin, 2002; Tocalli-Beller & Swain, 2005). They found that students noticed reformulations more than other types of feedback, and that reformulations led students to reflect on language errors more often. In a study by Storch and Wigglesworth (2010) that compared reformulations vs editing symbols, the authors found that editing symbols as feedback were incorporated, deliberated on, and revised more often by the students than reformulations. There are only few studies that have investigated the effectiveness of computer-mediated means of providing and delivering feedback. According to Bitchener and Storch (2016), these studies are not designed well (not having a control group or having pre, post, or delayed posttests) or do not frame their research in SCT or any other theories when discussing their findings. For instance, Yeh and Lo (2009) compared an online corrective feedback system to feedback printed on copies of student writings. Although the authors found that the experimental group performed better in identifying errors, there were no pretests or analyses of new writing tasks.

For studies that examine the behaviour of those who give and receive feedback, activity theory provides a useful framework to interpret human behaviour as an activity driven and defined by motives and realized by goal-directed actions (Bitchener & Storch, 2016). These studies fall broadly into two categories: students" response to teachers' WCF and teacher WCF practices. First, studies that attempt to explain students" response to feedback found that feedback provided may be noticed and revisions made; however,

feedback may not be incorporated in new pieces of writing due to feedback not being understood, which did not lead to writing development as seen in new pieces of writing compared to revisions (e.g. Storch & Wigglesworth, 2010). Students may not ask for clarification because doing so may be seen as challenging the teacher (Zhao, 2010), and students with lower proficiency were less interested in WCF than those with higher proficiency (e.g. Lee, 2008a). Second, studies on teachers' WCF practices show that teachers beliefs about WCF did not converge with actual practices in terms of quantity and type of feedback provided (e.g. Alshahrani & Storch, 2014; Lee, 2009), and the type of WCF given does not meet students' needs or preference (e.g. Lee, 2017). Bitchener and Storch (2016) point out that the number of studies on WCF from the SCT perspective are few and mixed because the WCF practices and students' responses are complex and multifaceted and due to the contextual and social nature of the SCT. As a result, the findings of these studies may not be generalizable to other contexts.

2.2.3 Methodological Challenges in Researching WCF

Although it seems that L2 and SLA research shows some agreement on the effectiveness of WCF, due to the differences in research design and approaches, there are questions of whether the two bodies of work on WCF can be compared, let alone provide practical pedagogical answers to writing instructors (Ferris, 2010). While, in general, SLA researchers have examined the acquisition of particular linguistic features in new pieces of writing using pre and posttest designs, L2 writing researchers have been more concerned with the development of writing as examined by students' revisions in response to teacher feedback in naturalistic classroom settings and employed comprehensive feedback to enhance ecological validity. For example, L2 writing studies such as Ashwell (2000), Fathman (1990), and Ferris and Roberts (2001) examined improvements in students' revisions of the same paper after receiving different types of WCF to assess the efficacy of one approach of WCF against another. In other words, L2 writing researchers focus on the effectiveness of different types of WCF and students' revision because results from students' revisions provide important evidence that helps teachers refine their practice (Ferris, 2010). However, critiques from SLA researchers

(e.g., Ellis et al., 2008; Sheen, 2007; Truscott, 1996) on examining students' revisions stem from the fact that examining revisions in drafts may not be measuring learning and that comprehensive feedback employed in the L2 studies makes it difficult to generalize their findings (Ferris, 2010). Ferris (2004) found that much of the literature on WCF has found dissimilar findings because the two veins of research are "not even asking the same questions to begin with" (p. 52).

L2 writing researchers contend that many SLA studies do not have ecological validity and do not reflect classroom practice. For instance, SLA studies tend to overgeneralize the correct use of specific functions of articles as accuracy gains (e.g., Bitchener, 2008; Bitchener & Knoch, 2009b; Bitchener & Knoch, 2009a; Bitchener & Knoch, 2010; Bitchener et al., 2005; Ellis et al., 2008; Sheen, 2007). The focus on a specific form as an indicator of accuracy gain is problematic because it is not clear if this focus would potentially hinder accuracy in other aspects, such as the overuse of articles, morphemes, verb tenses, sentence structures, and so forth. As Xu (2009) has stated, since some studies did not report any information about a change in the overall grammatical errors made by students before and after the treatment, students could have been more accurate with the target structures but made more mistakes in other forms since their attention was consciously directed to specific forms. For example, in the studies by Ellis et al. (2008) and Bitchener (2008), students knew which error types were being focused on in the research and may have focused more on the target features in subsequent writings. Therefore, the findings of the research have shown a more significant improvement of accuracy than if the students were not aware of the target features (Xu, 2009).

For SLA researchers, a short-term improvement in accuracy gains is not sufficient. For them, the salient aspect of WCF is whether learners can sustain this improvement. For instance, Ferris and Roberts (2001) found that students improved their accuracy with the teacher's help, but this gives no evidence that feedback has lasting effects on improvements in accuracy. Thus, SLA researchers contend that delayed posttests and more longitudinal studies are necessary to trace the development of accuracy over time.

To measure the change in magnitude of accuracy, SLA researchers argue that errors in the initial text must be compared to a new piece of writing after treatment. However, in many L2 feedback studies that examined comprehensive or unfocused feedback, only overall improvement in accuracy was measured regardless of whether the learners received feedback for that particular item or not. Thus, SLA researchers state that to precisely measure changes in accuracy in response to WCF, researchers would need to trace each type of error that received feedback (Guénette, 2007). Of course, L2 researchers assert that measuring each change goes against ecological validity in the classroom.

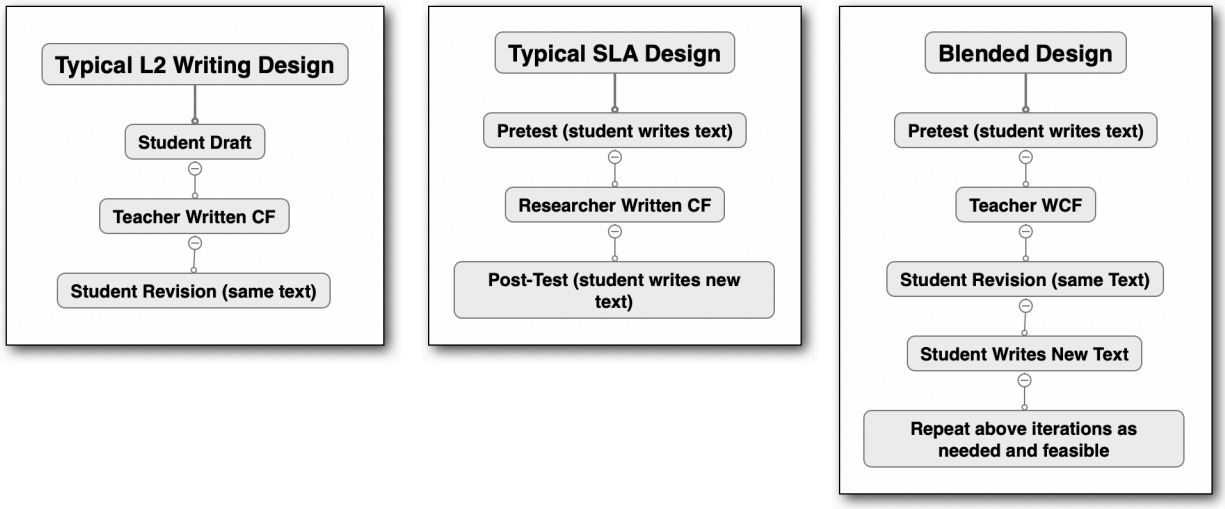
Moreover, in many L2 writing research studies, the design did not include comparison groups due to naturalistic classroom settings. These studies have relied on absolute gains made by groups receiving feedback. However, as Truscott (2007) has noted, "in the absence of a control group, one cannot determine whether observed gains resulted from correction or from other factors" (p. 263). Truscott (1996) has established that researchers must "compare the writing of students who have received grammar correction over a period of time with that of students who have not" (p. 329). SLA researchers affirm that research designs should include comparison and experimental groups in similar writing conditions and instructional contexts. In a meta-review of the topic by Ferris (2004), she found that very few studies of error correction in L2 writing actually "compare the writing of students who have received grammar correction over a period of time with that of students who have not" (p. 51). For example, studies by Fazio (2001), Kepner (1991), Lalande (1982), and Robb et al. (1986) did not have comparison groups. Therefore, because of the absence of a no-feedback comparison group, it is difficult to support the claim that error feedback is the primary reason and that time-on-task is not responsible for the progress measured over time (Ferris, 2011). In addition, in experimental designs, having a comparison group is not sufficient: researchers need to measure the pre-treatment differences to determine the magnitude and direction of the treatment. For example, Ellis et al. (2008) and Bitchener (2008) observed that in each study, the comparison group was noticeably weaker, but this was not considered in the analysis of performance improvement. Therefore, varied "accuracy performance after the treatment

between the groups, significant or insignificant, might not be pinned down to the treatment since the groups were not comparable to begin with" (Xu, 2009, p. 271). Fathman (1990) measured the pre-treatment differences but used a holistic scale. Guénette (2007) has suggested that although holistic scales are widely used to assess students, they might not be "fine-grained" enough for research purposes.

Lastly, due to studies being situated in classrooms, in many L2 writing studies, design parameters did not remain constant when comparing different types of feedback. For example, in a study by Lalande (1982), the two groups had different pedagogical activities, but how this may have affected the experiment was not investigated. If students in the comparison and experimental groups are engaged in different classroom activities, feedback effects are difficult to isolate. Moreover, in some studies, data gathering procedures made it challenging to identify the effects of feedback. For example, in some studies, the writing was done at home (e.g., Ashwell, 2000; Chandler, 2003; Sheppard, 1992). Thus, the time spent on the task and whether additional assistance was available were difficult to determine. In addition, in many early studies, the type of feedback and the number of feedback varied. Ashwell (2000) and Semke (1984) combined feedback on content and form, while Chandler (2003) and Robb et al. (1986) gave feedback only on language use. Furthermore, Chandler (2003) gave continuous feedback over a period, while Fathman (1990) and Ferris and Roberts (2001) gave feedback only on a single piece of writing. In addition, in all instances, how the feedback was delivered varied or was not specified. It cannot be assumed that all types of feedback on form are equal (Guénette, 2007). Lastly, varying incentives for writing may have affected students' motivation for both types of studies. For instance, some studies specifically gave scores for grammar (e.g., Robb et al., 1986), while others used assignments that were not graded or where accuracy was not the focus ((e.g., Fazio, 2001; Kepner, 1991; Semke, 1984). In some studies, participants were informed that they were being graded; this might have played a role in their motivation to pay more attention to form or discouraged students from writing more complex structures for fear of losing points (Guénette, 2007).

To help remediate the two different approaches to research designs, Ferris (2010) proposed a blended design that incorporates both types of methods from the L2 and SLA design (see Figure 2.1).

FIGURE 2.1
Possible Blended Design for Future Research



Note: Writing Research and Written Corrective Feedback. Adapted from "Second language writing research and written corrective feedback in SLA: Intersections and practical applications," by D. R. Ferris, 2010, Studies in Second Language Acquisition, 32, p. 195.

The proposed design incorporates examining both students' revisions and new pieces of writing to examine if WCF has an impact on revision and leads to the acquisition of correct forms. In addition, to have higher ecological validity, many teacher researchers argue that students' responses to WCF cannot be captured in studies where only a few target features are examined (Lee, 2008a; Mao & Lee, 2020) and does not reflect students' desire for comprehensive feedback (Lee, 2017).

In conclusion, although there has been much research on the effectiveness of WCF, many studies have been incomparable because of inconsistencies in design due to their SLA and L2 writing orientations (Ferris, 2004). However, a blended design that incorporates features of both L2 and SLA designs can be utilized for more robust research. Comprehensive feedback may reflect a more authentic classroom process where the goal is to improve all aspects of writing and not just a few targeted features. However, as

discussed above, many teachers do not have the resources to provide detailed feedback on students' writing, and AWE can help address this problem, as well as provide interactive feedback to support learning. Chapter 3 describes how the issues and recommendations discussed above informed the design of the current study. The next section discusses the literature on the role of AWE in the classroom, issues raised against its use, and previous research on the use of AWE to provide feedback.

2.3 Role of AWE / AES in the Classroom

Originally, Automatic Writing Evaluation (AWE) was primarily used in high-stakes testing situations. To solve the problems of marking consistency, speed, and reliability of human raters, Automatic Essay Scoring (AES) has been emulating "value judgments that human readers make when they read student writing in the context of large-scale assessment" (Herrington & Moran, 2001, p. 482). The AES scoring engine is based on dozens of features designed to measure specific aspects of essay quality. These features are derived using linguistic analysis, empirical modelling using statistical techniques, and a combination of both (Shermis & Hamner, 2012). The scoring engine provides a score based on a model of what human raters consider desirable (Attali & Burstein, 2006).

The technology originally developed for standardized testing is now being marked and developed for classroom use for both assessment and feedback. Grimes and Warschauer (2010) distinguished between the scoring engine (AES) and AWE because "they serve different purposes. AES is best known in the context of high-stakes testing; AWE is used for lower-stakes writing instruction" (p. 5). To further enhance AWE for the classroom, developers have extended the toolsets of AWE with a variety of resources such as integrated dictionaries, thesauri, portfolio system, a writer's handbook, graphical pre-writing tools, and student history repositories (Link et al., 2014). Early investigation has shown that computer-mediated writing instruction helped students write longer and be more sophisticated in their lexis and syntax (Schroeder et al., 2008). AWE systems should not be confused with online grammar checkers such as Grammarly

(www.grammarly.com) and LanguageTool (languagetool.org). These grammar checkers miss some critical features of AWE systems. To elaborate, although some grammar checkers can provide instantaneous feedback and metalinguistic explanations of some grammatical mistakes, they cannot be moderated by the teacher, do not evaluate writing quality, and do not include any portfolio and class management tools. Their focus is on correcting grammatical mistakes rather than writing development. Therefore, they are not discussed in this chapter (Ghufron, 2019; Nova & Lukmana, 2018; O"Neill & Russell, 2019; Park & Yang, 2020).

Currently, the three most commonly used tools in the classroom for writing development are Criterion, from a subsidiary of the Educational Testing Service (ETS), MY Access! of Vantage Learning, Inc., and Intelligent Essay Assessor of Pearson Knowledge Technologies (Warschauer & Ware, 2006). Common to all AWE programs for students is the provision of multiple opportunities for revising; the AWE systems generate immediate feedback on global writing skills and language use; and recent versions of these systems include other tools for effective writing such as model essays, scoring rubrics, graphic organizers, dictionaries, and thesauri (Stevenson & Phakiti, 2014). In this study, Criterion was chosen as the AWE system because of it includes a range of features, its widespread use in the writing classroom, and its convenience. Section 3.4.1 describes Criterion and its features in detail.

The use of AWE in L2 classrooms has given rise to a new term: automated written corrective feedback (AWCF) (Ranalli, 2018). AWE and AWCF allow students to "reflect on their errors and simultaneously track their effect on their overall performance, thus facilitating self-assessment, self-tutoring, and self-improvement through reflective use of feedback" (Yannakoudakis et al., 2018, p. 252). However, AWE feedback is currently not tailored to students" ZPD or tailored to the individual.

The distinction between direct and indirect written feedback may not be well-suited to AWCF because all of the current AWE systems' feedback results in some form of error

categorization and "may not be capable of generating specific suggestions for remedial action, depending on the error type and algorithm" (Ranalli, 2018, p. 3). Ranalli et al. (2017) proposed an alternative to the direct/indirect distinction: generic vs. specific feedback. In generic feedback, the same message appears when a category of error is detected and offer no specific prescription for remedies is offered. For instance, Criterion provides the same message when it detects a fragment: "This sentence may be a fragment. Proofread it to be sure that it has at least one independent clause with a complete subject and predicate". Specific feedback, conversely, incorporates some component of the text to give a recommendation. For instance, when Criterion detects a 'confused' words error, it incorporates the original text in its feedback, for example, "you have used 'a' in this sentence. You may need to use 'an' instead" (Ranalli, 2018, p. 3). Both general and specific feedback may engage students in guided learning and problem-solving as with indirect WCF (Lalande, 1982) because they may enable students to reflect upon their existing knowledge (Bitchener & Ferris, 2012), which may help in internalizing learning.

Initially, the development of automated scoring and feedback systems occurred primarily outside language learning and assessment with little direct influence from current language learning or testing theories (Xi, 2010, p. 219). Also, the AWE systems themselves were created with L1 users of English in mind (Burstein et al., 2004), which has resulted in most of the research being done on L1 learners of English (Grimes & Warschauer, 2010). However, as noted in the special issue of Language Testing on Automated Scoring and feedback systems in 2010, AWE is gaining popularity in the L2 context. To support this new field of research, Weigle (2013b) noted that "to evaluate the validity of automated scoring systems for these students, it is important to understand something about what it means to write in a second language, and how language proficiency and writing ability interact" (p. 87). Warschauer and Ware (2006) remarked that most of the studies on AWE in the classroom had been sponsored by AWE developers and that much of the research has been only presented at conferences; thus, "research conducted to date should be considered with a highly critical eye" (p. 7). As

teachers and institutions turn to technology to remedy the bottleneck of time constraints in the classroom in giving feedback, there needs to be research-based evidence for how best to use technology to provide feedback. Teachers and institutions "need a critically informed, empirically based inquiry that makes explicit both how, specifically, electronic feedback was used, as well as what criteria were used to evaluate its effectiveness" (Ware & Warschauer, 2006, p. 109). However, current literature does not frame research on AWE feedback in the classroom from either the cognitive or SCT perspectives. The following section explores the current limitations of AWE in the classroom.

2.3.1 Previous Research on the Use of AWE to Provide Feedback

There is evidence from classroom research showing that AWCF can improve the quality of students" writing (Stevenson & Phakiti, 2014). However, many researchers in the field are unsure of how successfully students can make use of AWCF for error correction because L2 classroom studies, so far, have been based on single-group designs (Chen & Cheng, 2008; Dikli & Bleyle, 2014; Lavolette et al., 2015; Wang et al., 2013). Moreover, while researchers like Kellogg et al. (2010) found that students" ability to reduce some types of errors after AWCF increased, the long-term effect of such feedback is unknown. The majority of previous research can be divided into three themes: teacher use of AWE, student motivation, and writing development.

Guichon and Cohen (2016) have suggested that teachers need to develop "semi-pedagogical competence" to facilitate L2 writing where they need to be aware of the semiotic affordances of tools and modes for the outcomes they want to achieve. Grimes and Warschauer (2010) observed that the teachers in their study found the integration of AWE in the classroom useful even though they found that the accuracy of the feedback was lacking. Nevertheless, the authors did not interview teachers to examine how the AWE system was integrated and why this was done. To fill this gap, a study by Chen and Cheng (2008) conducted interviews with the teachers on the issues of the integration of AWE and the use of scores and feedback. However, due to the lack of observational notes, how AWE was integrated in the classroom was not fully explored. In contrast, Link

et al. (2014) examined how ESL teachers implement AWE in the classroom and their perceptions of the experience of using AWE. The study found that teachers observed the benefit of AWE in increasing students" metalinguistic ability and reducing teacher workload but found that AWE was ineffective in providing necessary and high-quality feedback. This may, however, have been due to how the AWE tool was integrated in the classroom. The authors found that lack of familiarity with the AWE tool had a direct impact on teachers" ability to integrate it into their classrooms. In addition, like previous studies, teachers found that AWE promoted student autonomy and motivation.

Other studies have shown that AWE facilitates more writing practice and increases students" extrinsic motivation to write and revise due to instantaneous feedback (Burstein et al., 2004; Grimes & Warschauer, 2010; Li et al., 2014; Link et al., 2014; Wang et al., 2013; Warschauer & Grimes, 2008). These studies show that the enhancement of student autonomy and motivation to write more has the potential for developing students" writing. In addition, the holistic score provided by the AWE system may foster greater extrinsic motivation through rudimentary gamification mechanics (Hanus & Fox, 2015). For example, most AWE suites have components seen in most games, including points, badges, progress bars/progression charts, performance graphs, and levels. In video games, the challenge of winning a level or beating an obstacle can keep participants working until they can earn that achievement. While gamification is often based on extrinsic motivation (rewards, badges, points, and superficial game-like mechanics, etc), affinity for gaming can be leveraged in non-game-related contexts (Deterding et al., 2011) because the integration of AWE would make feedback/interaction immediate with the number of errors analogous to points and holistic scores to badges in games. However, this form of motivation may be less beneficial because students are more focused on scores than writing improvement and mastery (Deci & Ryan, 1985).

Like previous research on WCF, although conflicting in results, studies on AWCF have demonstrated improvements in accuracy. This efficacy may be due to AWCF being interactive, unlike traditional WCF. Although AWE feedback can be used summatively

after the draft is finished, the writing tools can be used during the writing process, providing feedback as the students write. Thus, there are more opportunities for the three processes involved in acquisition: internalization, modification, and consolidation (Williams, 2012). Shintani (2016) has written that when a student can revise the error immediately (modification), the student may have more opportunities to produce the same structures in later sentences (consolidation). With enough practice, this structure may be internalized. For instance, in a study that explored the impact of using AWE in the classroom in Taiwan by Wang et al. (2013), the authors found that the experimental group outperformed the comparison group in accuracy and enhanced their autonomy in the writing acquisition process. Likewise, Grimes (2008) found that students who used AWE wrote longer texts with fewer errors. In addition, Moseley (2006) found that the integration of AWE in the classroom led to positive changes in student perception of writing as a recursive process rather than a linear process reaffirming the concepts of Assessment for Learning and Assessment as Learning.

2.3.2 Critiques of AWE Feedback in the Classroom

The utility of AWE in EAP classrooms is tempting for teachers, but without evidence of its effectiveness, there would not be a broad adoption. At the same time, most AWE systems have been supported by favourable validity evidence based on the consistency and agreement between the automated system and human raters (Grimes & Warschauer, 2010; Weigle, 2013b). However, the potential instructional benefit in the L2 classroom of AWE remains unexamined. Some researchers believe that the high correlation and accordance between AWE scores and human ratings are an insufficient condition for the validity of score use (Li et al., 2015; Shermis & Burstein, 2003) because the human rater may be using criteria other than the ones in the rubric to make their decisions (Weir, 2005). Moreover, critics have expressed concerns over the accuracy of the scoring engine, lack of individualized feedback, lack of feedback on content and form, and the lack of relevance of the features evaluated by AWE to the more interactive qualities of writing, particularly the social and communicative dimensions (Ericsson & Haswell, 2006).

Much of the literature cites the inaccuracy of AWE. For instance, a recent study by Lavolette et al. (2015) examined a popular AWE system, Criterion. They coded all of the error codes that Criterion produced as correct, wrong, and no errors. They "determined that 1159 of the error codes (75%) were correct, 208 (14%) correctly identified an error but miscoded it, and 173 (11%) were for structures that were already correct" (p. 58). In addition, Criterion missed at least 46% of the errors. Because of this level of accuracy, Ranalli (2018) has conjectured whether this may influence students' willingness to use the resulting feedback.

However, a growing body of literature has shown that teacher feedback on error has been incomplete and inaccurate too (Cohen & Cavalcanti, 1990; Cohen & Robbins, 1976; Truscott, 1996; Zamel, 1985). Ferris (2011) has noted that in studies of WCF in L2 writing, "the teacher variable is usually either ignored or removed altogether from the research design" (p. 24). She also stated that even in studies where teacher feedback is analyzed, the feedback is seen as accurate, comprehensive, and consistent, and when more than one teacher is involved, the researchers presume that the teachers gave feedback in the same way. Thus, the usage of AWE feedback resolves many of these problems even with its low accuracy. Furthermore, while the instances of incorrect errors may be high, the feedback may be useful for students. For instance, "for advanced students with metalinguistic knowledge, the incorrect codes might still be highly beneficial if they promote noticing" (Lavolette et al., 2015, p. 64).

Research indicates that feedback of all errors may be overwhelming for L2 learners (Ellis et al., 2008). This calls for the reconfiguration of students' attitudes toward feedback. When students receive teacher feedback, they do not question it; however, for AWE feedback, due to its inaccuracies, the student must evaluate the error codes for correctness. Ferris (1995b) has advocated that students learn to self-edit and, with AWE feedback, students may be more cognizant of their evaluation process. Grimes and Warschauer (2010) proposed that "erroneous feedback is more frequently maleducative

when it is presented as authoritative" (p. 31). Thus, teachers and students should understand the limitations of AWE feedback. Lastly, because of these inconsistencies in the scores, teachers only had neutral or low trust in AWE. For instance, Li et al. (2014) found that teachers trusted lower scores more than higher scores from AWE.

Another criticism of AWCF is that due to limitations in technology, it is a one-size-fits-all software program. AWE suites are not designed to differentiate among users with different proficiency Levels or backgrounds. The AWCF that AWE suites provide for a run-on sentence error, for example, "is the same whether the user is an experienced writer whose L1 is English or an L2 learner enrolled in a developmental writing course" (Ranalli, 2018, p. 04). In addition, unlike previous studies of WCF, AWCF is comprehensive and is not sensitive to whether the error is treatable or untreatable. Thus, AWCF may not be valuable for all L2 learners, especially for those with lower competencies. These learners may not notice the errors with understanding, which is a necessity for uptake and L2 acquisition (Bitchener & Storch, 2016). In addition, because the error detection of AWE only generates fixed metalinguistic feedback for a particular error, the feedback does not account for individual differences of students in terms of their ZPD, knowledge of grammar rules, or background knowledge. In other words, the students may notice an error because it is highlighted by the AWE system, but the metalinguistic explanation may not lead to understanding if the students lack foundational knowledge. Furthermore, in studies that examined students' impression of AWE, students believed that AWCF was only useful for the first draft and was less useful in subsequent drafts because the feedback was repeated and thus ineffectual (Yang, 2004; Yu & Yeh, 2003).

Another problem is that holistic scores provided to teachers and students from the AWE do not measure content directly. Weigle (2013b) attested that "while [automatic scoring systems are] very consistent across prompts in total score, [they are] less consistent and thus less generalizable in terms of the specific features measured, particularly in the areas where second language writers may differ from first language writers" (p. 96). Thus, due

to what the AES engine is programed to consider as good writing, it may devalue creativity, risk taking, and other forms of writing. The scores provided by AWE may exercise power over what students believe is good writing if students do not understand the limitations of the scoring mechanisms and engage uncritically with it (Jason, 2020). For instance, the scoring may privilege certain organization strategies because that is what the algorithm sees as most optimal (Wang, 2015).

While validity, the approach to writing, and limitations of scoring accuracy and individualized feedback are essential, the most significant criticism of AES and AWE has been ethical concerns. The use of AWE in the classroom may distort students' notion of 'good' writing by privileging the types of writing that the AWE can give feedback on (Grimes & Warschauer, 2010; Herrington & Moran, 2001; Weigle, 2010; Weigle, 2013b). Another concern is that the use of AWE may diminish the role of the teacher and thereby reducing the human dimension in writing, which is against the fundamental purpose of writing. A position statement of the Conference on College Composition and Communication (CCCC) in the U.S. states:

> Writing-to-a-machine violates the essentially social nature of writing: we write to others for social purposes. If a student's first writing experience at an institution is writing to a machine, for instance, this sends a message: writing at this institution is not valued as human communication—and this, in turn, reduces the validity of the assessment (Communication, 2004, para. 12).

In addition, Canale and Swain (1980) asserted that the demonstration of linguistic mastery is a necessary but not sufficient condition for inferring communicative language ability on the part of the learner. The learners must be able "to respond to genuine communicative needs in realistic second language situations … not only with respect to classroom activities but to testing as well" (Canale & Swain, 1980, p. 27). This is currently the limitations of AWE feedback; it currently analyzes only surface and static features, and it cannot identify communicative intent, evaluate argument quality, and

most importantly, AWE systems decontextualize writing, depriving it of the social and communicative dimensions and eliminating the value of human audiences in real-world contexts (Communication, 2004).

Additionally, AWE feedback in the classroom for L1 English students has been criticized for promoting primarily a formalist approach to writing, in which writing is viewed as merely being the "mastery of a set of subskills" (Hyland & Hyland, 2006, p. 95) because of AWE's reliance on cognitivist models. Also, it can be argued that AWE models promote efficiency models of education, where efficiency is of foremost importance to pass standardized tests rather than learning through creative autonomy of students and teacher agency (Morgan, 2016). Comments generated by AWE have been reputed to place too much emphasis on the surface features of writing, such as grammatical correctness; therefore, student revisions in drafts were "overwhelmingly mechanical, primarily in spelling and grammar, rather than in organization" (Grimes & Warschauer, 2010, p. 8). In previous studies, L2 learners were dissatisfied with AWE's feedback system because they did not provide specific feedback on content and rhetorical aspects of their writing (Wang, 2015). Nevertheless, despite these criticisms, AWE may still benefit L2 English learners, where system limitations and problems are accounted for, because these technical features must be reviewed and mastered. The fact that AWE feedback and scoring systems focus primarily on mechanical aspects has some critics concerned that the promotion of AWE may lead to the automation of writing instruction (Warschauer & Grimes, 2008; Weigle, 2013b). The administration's promotion of giving feedback via technology may reduce teachers' autonomy, independence, and control over their work, ultimately leading to teachers becoming redundant (Iskander et al., 2010). However, although there is much discussion of AI and the erosion of teachers' roles and professional roles and students becoming more deskilled, functional, and technocratic, there is currently very little empirical research on the impact AWE and automation on L2 teaching and learning.

2.3.3 Methodological Challenges in Researching AWE Feedback

Some of the studies reviewed above have suggested that the use of AWE is justified if it is incorporated in the classroom using an approach that promotes thinking and critical reflection (Chen & Cheng, 2008; Dikli & Bleyle, 2014; Grimes & Warschauer, 2010). Other studies have noted that how AWE is integrated into instruction influences its use by the students (Chen & Cheng, 2008; Li et al., 2015; Ranalli et al., 2017). However, these studies had small sample sizes, which makes it difficult to generalize findings, failed to alleviate ethical concerns of integrating AWCF in the classroom, and lacked standardized integration of AWE in the classroom.

Most of the studies, so far, have had small sample sizes (Chen & Chang, 2008, n=53; Dikli, & Bleyle, 2014, n = 180; Li et al., 2015, n=67; Liu & Kunan, 2016, n=163). To generalize findings, longitudinal studies with a large sample are recommended to examine the development of writing proficiency due to AWE feedback. Also, although the results of previous studies have revealed improvements between drafts, they did not specifically examine improvements of writing proficiency of L2 in respect to students' L1. In addition, future studies need to be conducted with diverse ESL populations and programs to examine whether ESL writing development and the impacts of AWE feedback could be differentiated or generalized across grades, a student's English proficiency level, and course curricula.

There are ethical concerns over the use of AWE in the classroom. Li et al. (2014, p. 77) described situations where the teacher may overly rely on the scores and feedback that AWE provides, or students may develop surface-level features to improve their AWE scores without improving other aspects of their writing. In addition, some have noted that using computers for grading may be dehumanizing. Also, the use of an AWE may be disadvantage students who may be less competent with technology (familiarity with technological artefacts, skill sets with computers, access to equipment, etc.) and may cause anxiety using new technology.

Moreover, no AWE system can understand the context or the deeper meaning of language

typical of quality writing, nor infer a student"s aspirations, degree of risk-taking, or attempts to challenge themselves creatively or stylistically. The use of AWE may send the wrong message to students: that surface-level features are more important than organization and content. To mitigate these concerns, upon implementation of AWE systems, teachers and students should be trained on the limitations and use of AWE tools as part of the learning environment and their limited support-function in the process of learning and formative feedback. It is critical for teachers to integrate these tools in a way that both reduces the significance of the analytic scores and abstains from their use for gate-keeping to reflect the classroom practice to ensure that these technologies are not uncritically assimilated. Likewise, students need to understand the limitations of the feedback, so they will not be discouraged by feedback and scores that may be different from their teachers' as some of the students did in previous studies (Dikli & Bleyle, 2014; Li et al., 2014; Liu & Kunnan, 2016).

There are various possible ways of combining AWE with teacher feedback and scaffolding AWE feedback (Stevenson & Phakiti, 2014). However, in many studies, the integration of AWE was not prescribed but left up to the individual teachers as to how they would integrate AWE in the classroom. Despite this incongruency, these studies did not explore how AWE was integrated. While Grimes and Warschauer (2010) and Link et al. (2014) found that teachers had a positive perception of integrating AWE in the classroom, there were no observational notes to confirm the teachers' integration of the tool (e.g., Chen & Cheng, 2008). Due to the lack of observational notes, how AWE was integrated in the classroom was not fully explored. As teachers' familiarity with AWE increases, the way they integrate AWE will most likely change. This lack of standardization of how AWE was integrated reduces the generalizability of the findings of these studies and may introduce extraneous variables.

In addition, the different incentives to use AWE tend to be correlated with the students' uptake using the tool. For example, if the assignment was graded, there was a greater usage of the tool, and even more so, if the teacher assigned a partial grade for the AWE's

holistic score (Dikli & Bleyle, 2014; Grimes & Warschauer, 2010; Li et al., 2014; Link et al., 2014; Liu & Kunnan, 2016; Stevenson & Phakiti, 2014). However, currently, there are no studies comparing the differences in how different incentives affect the usage of AWE by the students. In the classroom, a critical integration of AWE may require a paradigm shift in the perspectives of teachers, students and administrators, teacher beliefs, attitudes and technology use, and administrative support to ensure new tools are not uncritically embraced and implemented in maximize efficiency. These can be of crucial importance because the "introduction of new technology often changes the broader ecology of the classroom, making comparisons to a comparison group difficult and the outcomes of technology can be so broad that they are difficult to assess" (Warschauer & Ware, 2006, p. 12).

Although AWE proponents may present AWE as a panacea for lack of time and resources in the L2 classroom, due to the many criticisms against AWE, it may introduce more problems than it solves: erroneous feedback, focusing on the mechanical aspects of writing and deteriorating professional conditions. The following section introduces the notion of hybrid feedback and some of the reasons why hybrid feedback would be more beneficial in the classroom and can address some of the problems discussed above.

2.4 Hybrid Feedback

While previous studies on WCF have shown that WCF may be conducive for writing development, differential success has also been noticed (e.g., Bitchener, 2012; Lee, 2008a). Ellis (2010b) has noted that these variations may be related to student engagement with WCF, which is mediated by multiple factors such as direct/indirect feedback, focused/unfocused feedback, students" L2 proficiency, and students" motivation, as well as the nature of the pedagogy and writing challenge itself. Recent research has suggested that student beliefs may change as they interact with the learning environment, and these changes in beliefs mediate their learning (Manchón, 2009). Student engagement with WCF is multi-faceted, including cognitive, behavioural,

sociocultural, and affective dimensions (Ellis, 2010b).

In addition, while WCF is closely tied to the work of teachers, much of the feedback studies have been conducted outside the classroom (Lee, 2014b). Some have suggested that the bulk of research, so far, has little pedagogical relevance and ecological validity (Storch, 2010) due to the clinical design of the studies. In comparison, much of SLA research examined the short and long-term effects of feedback on only one or two language features and lacks ecological validity (Bitchener & Knoch, 2015; Chapelle & Sauro, 2017; Kang & Han, 2015; Storch, 2010), and pedagogically valid research should focus on the potential of comprehensive corrective feedback that combines both higher and lower-level concerns (Van Beuningen, 2010). Thus, classroom-focused research should be concerned with the writing as a whole rather than the application of one or two grammar forms. Further, researchers such as Hyland (2010) have called for research that focuses on "feedback within the whole context of learning and on the learner"s role in interpreting and using feedback" (p. 181). As proposed by Han and Hyland (2015), WCF helps to generate revisions of writing and to internalize target structures; as noted previously in the ESL classroom, there is little resource to encourage the writing process, and "target structures" themselves may be theorized as abstract technical sub-skills disconnected from social, communicative, and aesthetic uses and locations in the world.

Furthermore, many AaL studies have raised concerns over students" reluctance to self-assess (Leach, 2012; Lee, 2016; Lee, 2017). Also, as explained above, L2 classroom teachers may not have the time to give constant and consistent feedback in a timely manner due to workload. To address the problem of the load WCF places upon teachers, Calfee et al. (2007) suggest that computers be adapted to "do some of the heavy lifting" (p. 284). However, as seen in the previous section, while the use of AWE in the classroom may help to solve these issues, it is not without criticisms and risks. Due to these criticisms, a growing number of scholars (e.g., Chen & Cheng, 2008; Li et al., 2015; Warschauer & Grimes, 2008; Zhang & Hyland, 2018) are advocating a hybrid approach to feedback that combines teacher feedback with AWCF. A hybrid feedback system may

help with greater learner autonomy, facilitate motivation, and potentially free teachers to devote more feedback to higher-level concerns. For instance, Weigle (2013a) has argued that because L2 learners have a greater need for feedback on sentence-level correctness, which AWE is adept at providing compared to feedback on higher-level concerns such as content, argumentation, and style, as well as a non-AI audience for communicative action, AWE could (if mobilized critically and reflectively with teachers' mediation) complement teacher feedback in L2 classrooms. If such technology is integrated in the classroom as a vital resource, and editing practices are encouraged, AWE may support teachers to build students' awareness about the importance of editing, more cognizant self-editing strategies, and metalinguistic feedback to improve students" writing by giving them feedback for AaL.

2.4.1 The Need for Hybrid Feedback

A hybrid feedback system has the potential to reduce issues associated with teacher feedback. The literature has recommended that feedback in the L2 writing classroom should be balanced and cover all dimensions of writing – content, argumentation, style, and language (Ferris, 2003; Hyland & Hyland, 2006; Zamel, 1985). However, a study by Lee (2008b) has shown that most teacher feedback provided by 26 Hong Kong secondary English teachers was on language form. Moreover, Biber et al. (2011) noted that feedback on both content and form is more effective than feedback on form alone. Thus, if feedback on the more rudimentary and technical aspects of form could be relegated to AWCF, teachers can spend more time giving feedback on content, argumentation, organization, and style to connect form with content, style, and genre. Also, a hybrid system may reduce problems of the current implementation of AWCF in the classroom. AES, the core of AWE systems, relies on features intended to measure the traits specified in holistic scoring models, such as the six-trait model offered by Spandel (2005), which has been a foundational guide for assessment in writing. However, Deane (2013) warns that AES may not match these models "due to the contrast between models focused on "text quality," measured in the end product, versus models focused on "writing skill," which is an attribute of the writer, not the text. In part, current limitations of AES

technology reflect the "differences in what kinds of features can readily be measured in the current state of natural language processing technology" (p. 12). This is a primary reason a hybrid approach may be needed, as well as understanding the differential functions of providing formative, immediate feedback for learning. Although research has found that AWE "can encourage learners to write more drafts, help them with noticing their errors, and draw their attention to linguistic features by providing metalinguistic explanations" (Mehrabi-Yazdi, 2018, p. 93), due to their limitations, Chen and Cheng (2008) have proposed that AWE feedback can be effective when it is combined with teachers' feedback. This may be because understanding writing requires the employment of cognitive and social, discursive and structural, temporal and historical, and linguistic and intertextual knowledge (Anson, 2006), which the current generation of AWE tools lack. The current rendition of AWCF is "not focused, graduated, contingent or dialogic" (Mehrabi-Yazdi, 2018, p. 95).

Compared to AWE feedback, the advantage of teacher feedback is the high level of personalization (Kakkonen et al., 2004). While AWE does give metalinguistic feedback, it is frequently generic. Therefore, while the metrics that are provided by AWE, such as sentence length or the use of passive constructions, may be useful, a teacher can better instruct students to recognize textual patterns in their writing to help develop students' metacognition to develop writing. In turn, this act can help situate the text in a social context because without social context, communicative purpose, or audience, writing "is not really a discourse; it is a bloodless, academic exercise" (Anson, 2006, p. 55). Consequently, having a teacher who knows the strengths and weaknesses of the student, and who provides feedback on content, style, critical thinking, and rhetorical knowledge may benefit the students' development in the writing process, as well as support their identity and role as a writer and language user. In addition, the teacher can give more personalized feedback by anticipating the needs of the students. Zhang and Hyland (2018) argued that although an experienced teacher can offer more comprehensive feedback on student writing, due to the heavy workload of giving feedback, AWE should be leveraged to maximize students' learning opportunities while ensuring that all

practitioners understand, as much as possible, the risks and limitations of implementing new tools.

While teacher feedback has been considered the gold standard of feedback studies, it is not without problems. Human judgment and feedback can be influenced by external factors such as being tired, halo effects (where human judgments of one aspect of writing is affected by their judgment of other aspects), stereotyping (where one's impressions about a particular group influence their judgment of individuals in that group), and other sources of inconsistency and bias (Zhang, 2013) (of course other biases may exist in the algorithms of AWE and AES because humans program the AI). In addition, there are often misalignments of instructors' beliefs and practices. Mao and Crosthwaite (2019) found that while instructors believe global issues such as argumentation and organization are more important for writing development, they spend more time on mechanical issues. Moreover, although instructors provide error codes to the students, it is not usually accompanied by metalinguistic explanations.

Therefore, combining critically informed uses of AWE feedback with instructor feedback may have more advantages than using either alone: the combination may garner greater student participation, be more systematic in treating errors, reduce cognitive load on teachers, and reinforce the supposition that writing is crafted in social contexts. First, active student participation and engagement are crucial if the language learning potential of written corrective feedback is fully exploited in tertiary contexts (Hyland, 2010). The sociocultural theory proposes that learners benefit most when prompted to self-correct and scaffold their attempts within their ZPD (Ellis, 2010b). However, teachers may not have the resources due to overwork and institutional priorities to give personalized feedback for each student's specific needs. In addition, researchers have found that WCF from teachers may negatively affect certain aspects of engagement, such as uptake. For example, Sheen (2008) found that language anxiety reduced students' engagement with WCF. However, studies in Computer Assisted Language Learning (CALL) have suggested that the use of technology may reduce anxiety, promote autonomy and increase

engagement (e.g., Chapelle et al., 2008; Peterson, 2010; Roed, 2003). Moreover, the affordances of hybrid feedback provide both immediate digital (interactive) feedback and delayed human (teacher) feedback, which may be more closely aligned to students' needs. In other words, because the writing-feedback loop can occur anywhere and at any time, more agentive and participatory learning may occur. In addition, because AWE has a discernible advantage over teacher feedback regarding the timeliness, convenience, multiple drafts, and learner autonomy (Chen & Cheng, 2008; Dikli, 2006), it would provide more opportunities to revise an essay multiple times at the students' own pace (Warschauer & Ware, 2006). Therefore, the number of drafts that a student produces, which equates to time-on-task, should not be limited. If the tool is positioned as a technology for assessment as learning, and students understand how to mobilize the tool for these purposes, the use of AWE can empower students with "the responsibility to revise their essays according to their own schedule, enjoying the autonomy offered by AWE feedback and, by capitalizing on the multiple revision opportunities to improve their drafts, and internalize language points in the revision process" (Zhang & Hyland, 2018, p. 100), which reflects and fosters the process-oriented approach to writing. This approach supports utilizing an AWE system critically without students feeling they are being "scored" or summarily judged by it. This view supports an understanding of how to use the tool in ways driven by student purposes and meta-cognitive strategies because students see the AWE system as something they can use, work with, reflect on, and not as a gatekeeper, judge, or testing apparatus.

Second, due to the sheer complexity of the issues involved in correcting errors and the varying classroom contexts, studies have shown that teachers' treatments of errors are inconsistent and imprecise (Zamel, 1985). Some studies have suggested that teachers are not capable of giving correct grammatical feedback (Lee, 2004; Truscott, 1996). Moreover, there can be a mismatch between the teachers' and students' goals. Scholars have observed that in many cases, teachers tried to control the feedback process due to the demands of the curriculum and institutional needs (Hyland & Hyland, 2001; Lee, 1997; Lee, 2004; Truscott, 1996); the teachers failed to consider the students' own goals –

relegating the students' learning process as secondary. However, if students can engage critically with the AWE feedback, in addition to teacher feedback, students may utilize the tool for their own learning goals.

Third, in addition to reducing time and cognitive load on teachers, Kakkonen et al. (2004) propose a semi-automatic or a hybrid approach that can aid the teachers in the following three ways: it can assist the teacher to grade the essays; it can support the student during the essay writing process; and, it can make the grading process more visible in the sense that some criteria for grading and feedback about the essay are available for the student (p. 458). However, as the integration of technology in the classroom is affected by prior beliefs, practices, and social-institutional settings, it is unclear if a hybrid feedback system will foster pedagogical approaches in the classroom. It is yet to be determined if the hybrid system supports the language competencies of all members of a learning context for multimodal and digital literacies.

Lastly, Herrington and Moran (2001) opined that writing becomes reduced and degraded as we write to machines. Studies show that in using AWE in the classroom, writing is framed as a product to evaluate student mastery of grammar, usage, and organization, and it is not modelled on providing meaningful feedback and lacks negotiation between the reader and the writer, or among peers (Wang & Brown, 2008; Ware & Warschauer, 2006). It also limits dynamic opportunities for collaborative writing projects on Wikimedia or other web-based media/sites where students are engaged in collaborative co-authorship and multimodal textual making (Thumlert et al., 2015), as well as collaborative inquiry and multimodal writing/creation involving diverse language learners in authentic multilingual/plurilingual contexts (Thumlert et al., 2018). Scholars have remarked that computers cannot replace interaction with teachers because ESL writers need instruction, modelling, and practice (Reid, 1994; Warschauer & Ware, 2006). If only AWE systems are used to give feedback, it can reinforce artificial, mechanistic, and formulaic writing driven by the algorithms giving the feedback (Wang et al., 2013). Implementing novel technology tools may also predetermine and limit the range of admissible learning

challenges and artefactual outputs (Thumlert et al., 2015). However, if the more mechanical aspects of writing are relegated to AWE by interactively giving feedback on linguistic microfeatures (Crossley et al., 2014), and more creative aspects are assessed and given feedback on by the teacher (Chen & Cheng, 2008; Grimes & Warschauer, 2010; Warschauer & Grimes, 2008; Warschauer & Ware, 2006), some of the criticism against AWE would be addressed. The opponents of using AWE in the classroom have voiced that the integration of AWE would lead to students writing noncreative essays (Communication, 2004; Ericsson & Haswell, 2006; Herrington & Stanley, 2012; Perelman, 2012). Having said that, a hybrid approach that critically blends both AWE and teacher feedback may be more inherently a part of a constructive, interactive, and formative learning process (Kakkonen et al., 2004), one that provides both immediate corrective feedback and human dialogical support and modelling while avoiding above mentioned problems and at the same time guaranteeing teacher professional identity.

2.5 Conclusion

Warschauer (2010) noted that the use of machine scoring appears to conflict with the goals of a sociocognitive approach to writing as elaborated above; however, the research reviewed here suggests that the impact of AWE in the classroom largely depends on how it is used, theorized, and positioned sociotechnically. As with many instructional strategies and innovations, the tools of technology provide the most beneficial results when integrated into a strong curriculum and when clearly matched to instructional purposes. Consequently, empirical studies should inform teachers about how to combine teacher and AWE feedback more effectively in the classroom.

There is a considerable variety of opinions in what constitutes purposeful writing; the underlying assumptions that teachers hold about literacy are integral to how technology is integrated as a resource into writing classrooms, and this is shaped by the teachers' beliefs about the software. As with any technological advancement, the potential exists to better support student learning or, alternatively, alienate students and teachers. In the case of

AWE, the potential for teachers to devote valuable time for more drafts and writing practice exists. Likewise, for the student, the immediacy and personalized feedback can offer more practice of writing and motivate them for further practice. The studies above propose that AWE could have a significant role in the EAP writing classroom in conjunction with teacher feedback in developing academic writing.

As the limitations of AWE become more evident in the literature, the integration of AWE and teacher feedback may be a more critical approach in the classroom (Zhang & Hyland, 2018). With the utilization of AWE feedback, teachers can adjust their feedback focus and allocate more time to give feedback aimed at the rhetorical development of student revisions and the metacognitive skills of the writing process. There is a need to examine how this hybrid feedback system affects student engagement in revision, student autonomy, and the writing process. Also, an increased understanding of how students and teachers engage with hybrid feedback and a greater awareness of how their attitudes towards hybrid feedback influence the conceptions of teaching and learning of writing is needed. This study addresses these questions, and the following chapter describes how these questions are addressed in the current study.

Chapter 3: Methods

Chapter 3 presents the methods used in the current study to examine the writing performance of students in an EAP class utilizing AWCF in addition to receiving teacher feedback. Specifically, this chapter provides details about the study's context, design, participants, and data collection and analysis procedures.

This study adopts a recommendation by Ferris (2010), as mentioned in Chapter 2 for studies on feedback to employ a quasi-experimental design with a comparison group and executed in the classroom using regular course assignments with a pretest, treatment, and delayed tests to examine the efficacy of hybrid feedback in a context of an EAP writing course. The design would compare the writing practices and change in language between an experimental group and a comparison group, which did not receive hybrid feedback. This study aimed to combine the quantitative features of SLA research and the qualitative characteristics of L2 writing research in what Ortega (2012) describes as L2 writing SLA interfaces.

The study is a classroom-based investigation to have greater pedagogical relevance for language teachers and greater ecological validity. The primary focus is on the effects of hybrid feedback on students' writing practices, language, and beliefs about writing compared to the comparison group. The writing tasks, apart from the delayed posttest, were fully integrated into the curriculum. By aligning writing tasks for the study with their in-class ongoing assessment tasks (The EAP program includes a diagnostic writing test in week one and three in-class essays afterwards), the solicited data was naturally occurring samples which are more valid evidence of student interlanguage development (Ellis & Barkhuizen, 2005). Ellis and Barkhuizen argued that timed essays written in examination provide samples of student data since an examination "constitutes a 'natural' context for students to use the L2, and data so obtained have not been designed for purposes of research" (p. 50). In this way, the feedback was given within the context of an instructional program, with authentic writing tasks, and where revision and the writing

process were meaningful because they were embedded in the curriculum and were reflected and reinforced in what is taught and emphasized in the class (Storch, 2010). Due to the intensive nature of the classroom context, the study was longitudinal with multiple treatment occasions.

3.1 Study Context

The study took place at an EAP program at a language institute at a university in Southern Ontario. The institute is housed within the School of Continuing Studies, and it is the largest English language institute of its kind in Canada. The classes are taught by teachers with a graduate degree in Applied Linguistics or other relevant majors such as Teaching English as a Second Language (TESL). The students in the program are mainly mainland Chinese who have finished high school and have conditional acceptance to various undergraduate majors; thus, they are highly motivated with a similar educational context and the same mother tongue. "This EAP program began with the first four weeks of the course with two parts a day for a total of 6 hours per day, five days a week for eight weeks. The first part was 4 hours in duration that focused on reading and writing skills, and the other was a 2-hour part specific to listening and speaking skills. The current study took place in the 4-hour reading and writing part of the course. In addition, in the reading and writing sessions, starting in week five, 2 hours on Tuesdays and Thursdays were reallocated to an elective segment until the beginning of week 8. This reduced the amount of time in the reading and writing part of the course from 20 hours per week to 16 for the three weeks of the course." The reading and writing class focuses on writing features such as content development, developing and organizing ideas, and grammatical and lexical accuracy and sophistication. The program's goal is to have students meet the university language requirement of IELTS band 6.5 by the time they matriculate to their degree programs.

There are usually between 15-17 students in a class. The students are randomly placed by an administrator into different class sections; although each class can have a mix of ages,

chosen majors, first languages (usually Mandarin or Cantonese), regions, and other demographics, the population of the program is relatively homogenous. The writing curriculum covers three types of essay structures: cause and effect, compare and contrast, and argumentative. Unlike the listening and reading assignments, teachers are free to choose the essay prompts for the writing assignments.

The program was selected for the study due to the writing proficiency level of the students, a classroom context that facilitates multiple treatment occasions, a writing curriculum that includes writing tasks that are similar to writing tasks that AWE systems use and can analyze and give feedback on, and the freedom to choose the bank of writing prompts from the AWE. As discussed in the previous chapter, AWCF's indirect/generic feedback may be more amenable for intermediate and advanced proficiency levels. Accordingly, the requirement of overall IELTS band 5.5 for entering the program is well suited for this study because the generic feedback that AWE provides may be more suited for students with higher proficiency. Against this backdrop, the current study introduced hybrid feedback in a quasi-experiment to investigate the effectiveness of such feedback.

3.2 Research Design

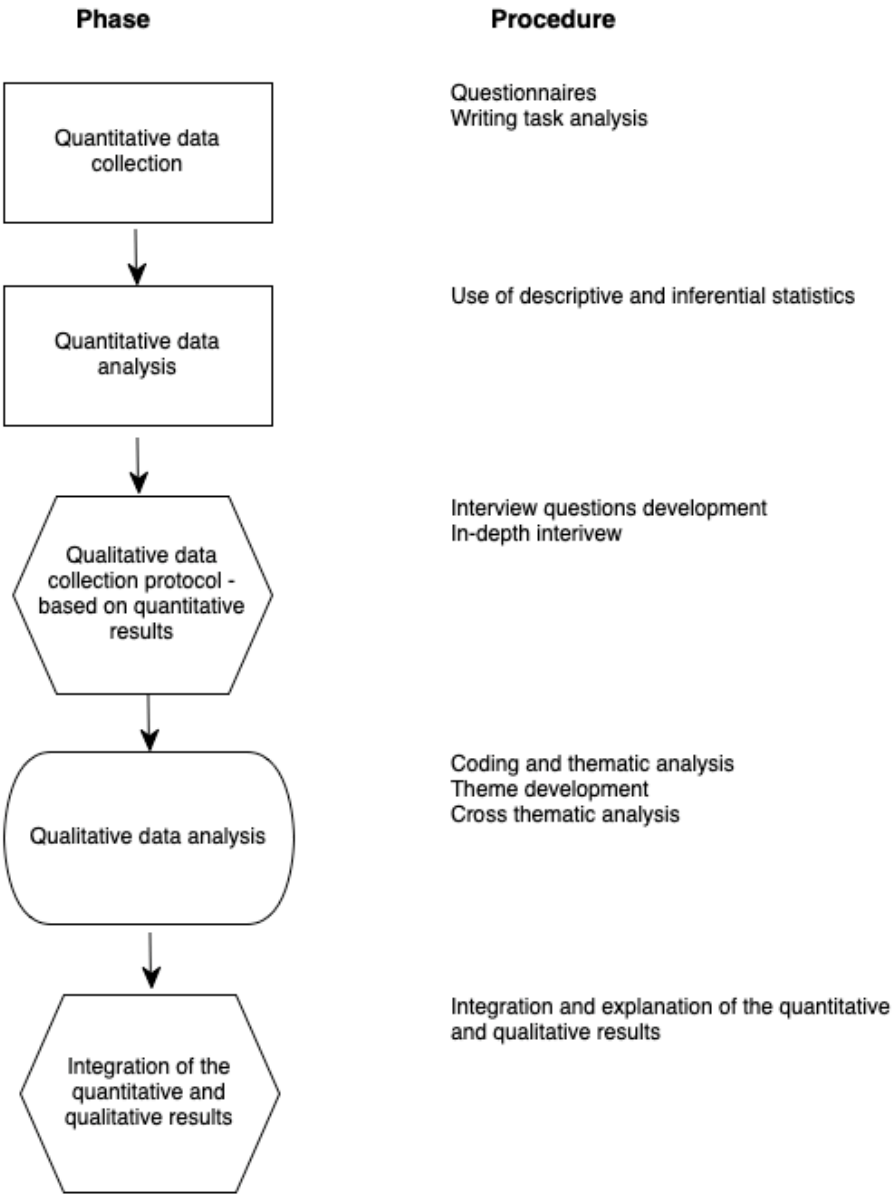
This study's design was quasi-experimental to compare two groups of students from two intact classes, one using hybrid feedback and a second similar one with only teacher feedback. In both groups, students received feedback and scores from the teacher. However, in the experimental group, before submitting to the teacher, students had the option of submitting their essays to an AWE system to receive AWCF until they were satisfied with the revisions. Once they were satisfied with the writing, the final version of the essay was submitted to the teacher. In both groups, students could use teacher feedback to improve their writing in their future essays. Both groups participated in pre, post, and delayed posttests to measure the differences in the impact of hybrid feedback and instructor-only feedback on writing improvement.

I took steps to reduce bias because I was the researcher and the teacher for both classes. To ensure that the decision to take part in the research was voluntary, I asked an administrator not involved in the study to meet with the students at the beginning of the semester and provide an informational session to reduce the power differential when recruiting. The administrator also told the participants that they could withdraw from the study at any stage if they wished to do so and that their participation and withdrawal from the study would not affect their scores in the class to ensure that the potential participants were free of undue influence and/or coercion. The administrator also informed the participants that the data would be destroyed for those who did not want to participate in the project. In addition, I collected all data but did not analyze them until after the final marks were submitted to ensure that there was no pressure on the students to participate in the study, and I would not know who was participating.

A mix of quantitative and qualitative methods was used in the study. This study examined if the combination of feedback resulted in differences between the two groups' approaches to writing, language, content, and organization of writing and explored how the different types of feedback affect students' writing practices. A quantitative-qualitative sequential triangulation mixed methods design was employed by collecting quantitative data first and then augmenting the quantitative results with in-depth qualitative data (Creswell & Clark, 2007). To answer the first research question, in the quantitative phase of the study, questionnaires on feedback and writing practices were administered to both groups pre- and post-treatment to see if there were any differences between and within groups. To answer the second research question, changes in the language, content, and organization of both groups' writing were analyzed for each writing occasion and for revisions between and within groups. To answer the third research question, questionnaire data for the experimental group's perception of hybrid feedback were analyzed. At the end of the treatment, a focus group interview was conducted with students in the experimental group as a follow-up to the quantitative stage to augment the quantitative results. In this explanatory follow-up phase, findings from quantitative data were used to develop focus group interview questions to clarify, supplement, and add to the

understanding of the students' views of hybrid AWE feedback and changes in their writing practice (See Figure 3.1). The sequential explanatory design allows for deeper insights to the context behind the statistical results (Field, 2018).

FIGURE 3.1
Quant-Qual Sequential Triangulation Mixed Methods Design



3.3 Participants

From the EAP program, 33 students from two intact classes participated in the study, with one class being the experimental group with 17 students and the other the comparison group with 16 students. All students in both classes volunteered for the study, and the recruiting procedure is detailed in section 3.6, Data Collection Procedure, below. The study used a non-probability sampling method (Teddlie & Tashakkori, 2010). Although random samples may be preferable for generalization to a larger population, it is not feasible in classroom-based studies. There were 16 students in the comparison class with ten males and six females with an average IELTS score of 5.47, with two students not having IELTS scores. The experimental group consisted of 17 students with nine males and eight females with an average IELTS score of 5.5, with one student not having an IELTS score. The IELTS writing scores for the experimental group were lower than the comparison group, 5.20 compared to 5.39. The comparison and experimental groups are comparable in terms of other characteristics, including age, years of learning English, and overall IELTS scores. Information about the students in both groups can be found in Table 3.1 below.

TABLE 3.1
Student Demographics

	Comparison Group		**Experimental Group**	
Total Participants	16		17	
Gender Distribution	10 males	6 females	9 males	8 females
	Mean	SD	Mean	SD
Age	19.13	1.54	19	0.94
Years studying English	9.69	2.02	10.06	2.88
IELTS Overall Score	5.50	0.34	5.47	0.22
IELTS Writing	5.39	0.35	5.20	0.41

3.4 Treatment: Hybrid Feedback

Hybrid feedback, as its name implies, is a hybridization of teacher and automated feedback. The students in the experimental group received a combination of AWE

feedback and teacher feedback, where the AWE feedback focused on lower-level aspects. In contrast, the instructor also gave WCF on lower-level aspects but focused more on higher-level aspects of writing and mediated AWCF in a follow-up oral conference. The comparison group received only teacher feedback following the same guidelines as teacher feedback for the experimental group.

3.4.1 AWE Feedback

Criterion, a web-based AWE tool developed by ETS, was used to provide AWECF in the study. Criterion was chosen from other available AWE tools due to the wide range of features it includes, its wide usage in the classroom, and its convenience. Criterion includes feedback on essays through a holistic score and trait feedback analysis, which gives a score for three aspects of writing: word choice; grammar, usage, and mechanics; and organization, development, and style, which mirrors the analytic rubric prescribed in the curriculum. Also, the essay prompts and types of essays offered by Criterion match the types used in the curriculum. In addition, the market reach and wide usage of Criterion mean that the results would be of more interest to more people; Criterion is widely used in K-12, ELL, and university settings. The scoring engine for Criterion is also used as the second marker for the TOEFL test, a widely accepted standardized English test for university English proficiency that is pertinent to the context of the EAP writing class. The web-based nature of Criterion meant that no IT support would be needed for installation and maintenance. Lastly, ETS was responsive to my questions and request to use Criterion in this study.

With Criterion, students receive feedback after the writing is completed by copying their text into the window. Criterion generates both surface-level feedback on mechanical and grammatical errors and content-level feedback, which provides generic feedback such as highlighting the first sentence of each body paragraph and asking if the following sentences support the topic sentence as is expected in a formalized academic writing appropriate to the EAP context. The participants were free to choose which feedback they used. The focus of AWE feedback is on lower-level aspects of writing. The Criterion

version used in the current research was 19.3.0, and the e-rater scoring engine version was 20.1.0. Criterion has two interdependent applications: (1) Critique Writing Analysis Tools, which detects errors in grammar, usage, and mechanics, identifies discourse elements and recognizes potentially undesirable elements of style; and (2) an automated essay scoring (AES) system - E-rater (Burstein et al., 2004). The trait feedback analysis gives feedback on grammar, mechanics, usage, style, organization, and development. For example, the grammatical errors that Criterion detects and provides feedback on include sentence fragments, missing commas, run-on sentences, subject-verb agreement errors, ill-formed verbs, pronoun errors, possessive errors, and wrong or missing words (Attali, 2004). Criterion gives metalinguistic explanations for each type of trait feedback and a summary of errors for each type. The suite also includes a "make a plan brainstorming app," which the student can use to plan essays; a "writer's handbook," which is a style and grammar guide; a spell checker; a portfolio option to keep all writing tasks organized; and a help guide for the suite. Criterion does not provide a way to turn off these features, nor does it provide a method of tracking their use. These features are similar to resources freely available on the web and in most word processors; therefore, students' use of these features is assumed to be part of the writing process and to not have affected the study findings.

Criterion gives feedback in the following five categories: Organization & Development, Grammar, Usage, Mechanics, and Style. The organization & development category highlights and gives generic feedback on Introductory Material, Thesis Statement, Topic Relationship & Technical Quality, Main Ideas, Supporting Ideas, Conclusion, and Transitional Words and Phrases. For example, Criterion highlights all first sentences of body paragraphs and gives the following feedback:

> Criterion has identified three or more main ideas in your essay. Do these ideas support the thesis statement of your essay? Do you use examples, explanations, and details to support and extend your main ideas? Does everything connect back to your thesis statement? Look in the Writer's

Handbook for ways to develop main ideas.

The style category highlights and gives generic feedback on Repetition of Words, Inappropriate Words or Phrases, Sentences Beginning with Coordinating Conjunctions, Short Sentences, Long Sentences, and Passive Voice. For example, for repetition of words, Criterion highlights frequently repeated words and gives the feedback, "You have repeated these words several times in your essay. Your essay will be stronger if you vary your word choice and substitute some other words instead. Ask your teacher for advice." For each area of suggestion that students click on, Criterion highlights the specific word or phrase. When students hover their mouse pointer on the highlight, a pop-up screen gives a further explanation, as in Figure 3.2 below. For form-focused feedback, Criterion breaks it down into the following three categories: grammatical errors, word usage errors, and errors in writing mechanics. Table 3.2 illustrates error categories generated by Criterion.

FIGURE 3.2
Example Screenshot of Criterion Feedback for Organization & Development - Transitional Words and Phrases.

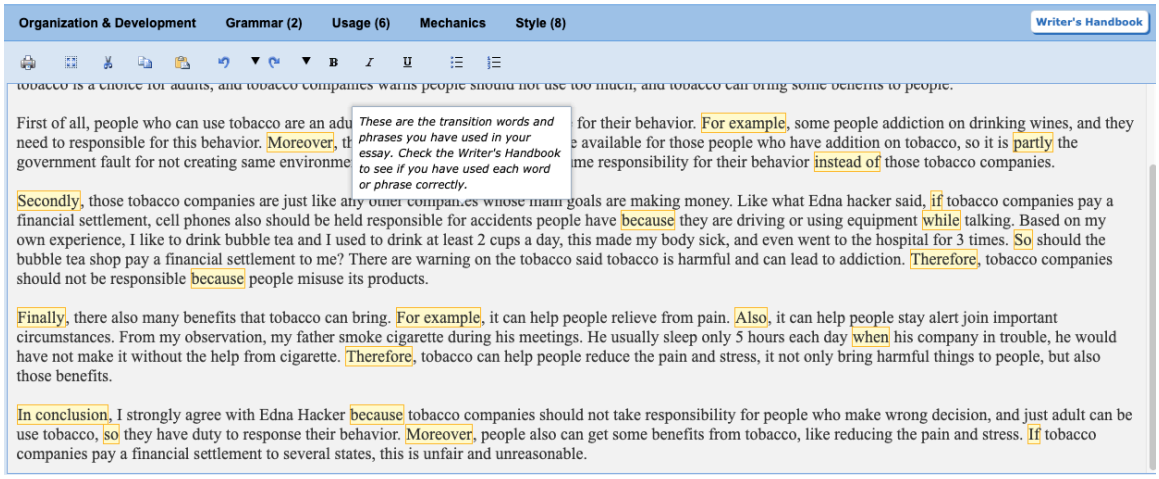


TABLE 3.2

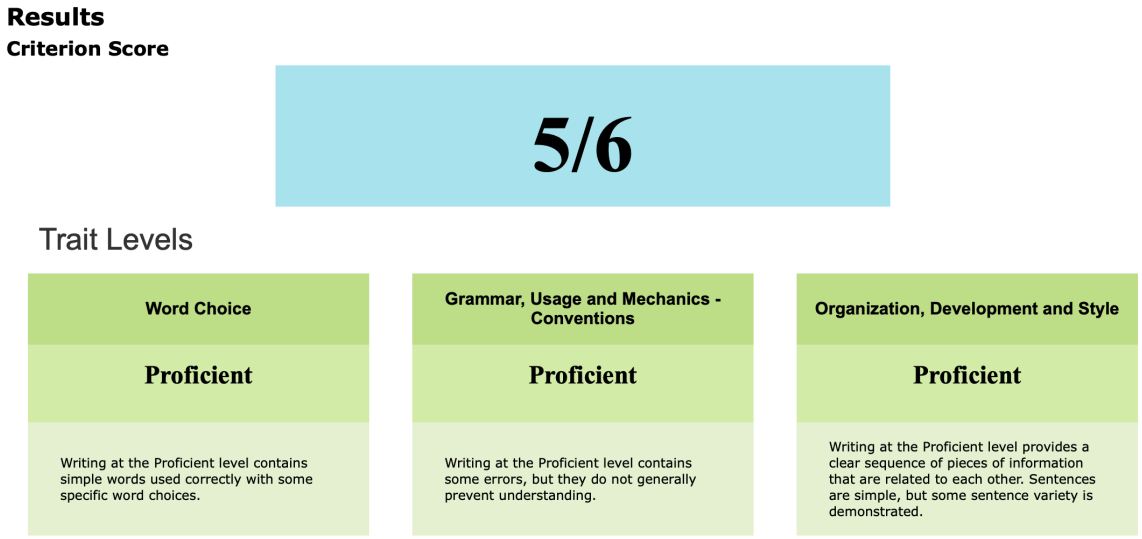
Criterion Trait Feedback for Version 11.1

Grammar Errors	**Usage errors**	**Mechanics errors**	**Style**	**Organization and development**
Wrong or missing word	Missing or extra article	Spelling	Repetition of words	Presence of a thesis
Ill-formed verbs	Wrong article	Capitalize proper nouns	Passive voice	Main points
Proofread This	Determiner-noun agreement	Missing initial capital letter in a sentence	Too many short sentences	Supporting ideas
Subject-verb agreement	Confused words	Missing question marks	Too many long sentences	Presence of a conclusion
Pronoun errors	Wrong form of word	Missing final punctuation	Sentences beginning with coordinating conjunctions	Transitional words and phrases
Garbled sentences	Faulty comparisons	Missing Comma	Sentence variety	
Fragments	Nonstandard word form	Missing Apostrophe	Inappropriate words or phrases	
Possessive errors	Preposition error	Hyphen error		
Run-ons	Negation error	Extra comma		
	Parts of speech	Fused words		
		Compound words		
		Duplicates words		

Note: Organization and Construct Coverage of e-rater v11.1. Adapted from Evaluation of e-rater for the GRE issue and argument prompts, by C. Ramineni et al., 2012, ETS Research Report, 12(02), p. 40.

At the beginning of the course, students logged in to the program with a class access ID and password that I provided to create their usernames and passwords. While teachers can create their own prompts, for which Criterion provides feedback but no scores, in this study, prompts from the Criterion database were used because studies have shown prompt-specific scoring and feedback is more accurate (see Chen et al., 2017). These prompts matched the discourse modes taught in the class: cause and effect, compare and contrast and argumentative. Students typed their responses directly in Criterion or copied and pasted them from Microsoft Word or Windows Notepad. When they submit an essay, the student receives a performance summary from Criterion that includes a holistic score, score summary information, trait scores, and feedback (see Figure 3.3 for an example).

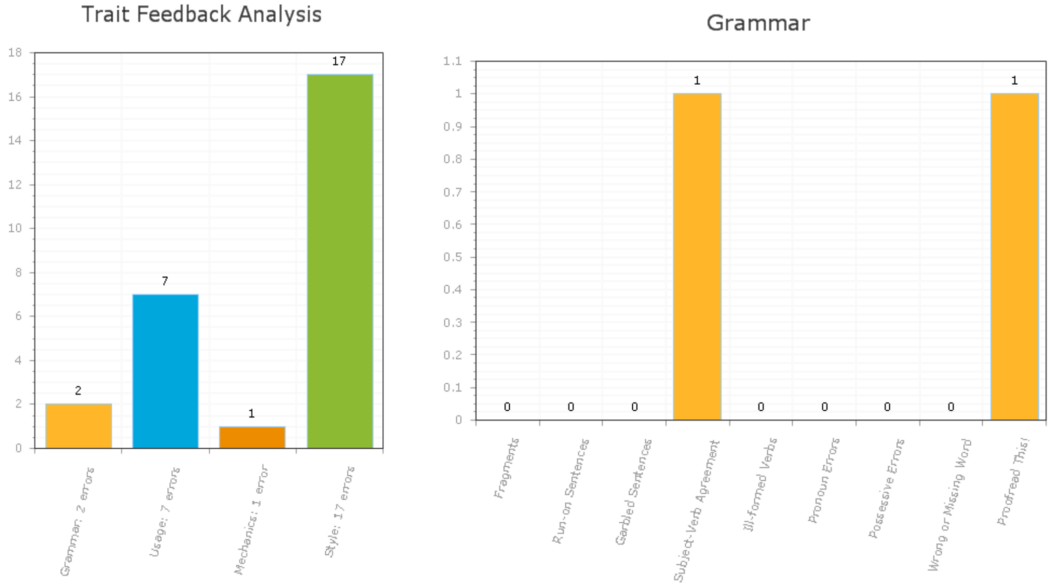
FIGURE 3.3
Example Screen Shot of Criterion Score and Trait Levels



The holistic score given by Criterion is given on a six-point scale, a score of six representing a high-quality paper (Ramineni & Williamson, 2013) (See Appendix A for the description of the ETS Criterion Score Guide) and analytic scores for word choice; grammar, usage, and mechanics; and organization, development, and style. The analytic scores are given on a three-point scale: developing, proficient, and advanced. The students could get more information about individual traits by clicking on the individual links to the trait feedback analysis screen. Criterion provides a graphical summary of mistakes in the various categories in the trait in the analysis screen (see Figure 3.4 for an example). If there are errors in the categories, students can click on them to see the location of the mistakes in the essay and roll over the mouse for metalinguistic explanations. For instance, for misuse of "there," the roll-over feedback would say, "you have used there in this sentence. You may need to use they"re instead." In the roll-over message, the students are provided with a link to the Writer"s Handbook, which provides more information on the error.

FIGURE 3.4

Example Screenshot of Graphical Summary of Mistakes



3.4.2 Teacher Feedback

Chapter 2 has described the complexity of feedback and how there is no consensus in the literature to guide teachers on a framework for giving feedback, except that teachers" feedback decision making needs to be situated in a specific classroom context because "feedback is deeply influenced by contextual issues such as students" characteristics, teacher beliefs, as well as the larger institutional context that governs teaching and learning" (Lee, 2017, p. 74). Researchers have noted that prescribing the types, the amount, and commenting styles of feedback lacks ecological validity because it does not reflect classroom practices or instruction at that instance (Lee, 2017; Polio, 2017). Therefore, I chose "best practices" of feedback that can be incorporated into the program"s EAP context. The following guidelines were used in providing teacher feedback for both the experimental and comparison groups in this study.

A balanced approach to feedback: Bitchener and Ferris (2012) have advocated teachers adopting a middle approach that combines focused WCF with comprehensive WCF. Focused WCF should focus on targeted areas of accuracy, and comprehensive WCF should focus on global content, organization, and development issues. In this study, comprehensive feedback was provided on the introduction paragraph and the last body

paragraph only following a suggestion by Evans et al. (2010) that comprehensive WCF should be done for selected paragraphs rather than entire texts. When commenting on content, argumentation and style, questions were used to give feedback on content, argumentation and style, and a written commentary was given. While written commentary can take several forms, including statements, imperatives, questions, and hedges (Ferris, 1997; Sugita, 2006), research suggests that questions are generally more desirable because they can enhance cognitive engagement and autonomy (Ferris, 2014), and for content, they can help students clarify their ideas (Nurmukhamedov & Kim, 2009).

Customizing feedback to individual student needs: It is essential for teachers to give feedback according to the needs of individual students (Evans et al., 2010; Ferris, 2006; Han & Hyland, 2015; Hyland, 1998; Lee, 2008a). Even though the students in this research context may have overall IELTS scores of at least 5.5, there may be a variety of proficiencies in writing. Therefore, it may be that weaker students may find hedges in teacher written commentary confusing (e.g., you may need to develop this point more). Lee (2017) has suggested that students may be better served by receiving more direct comments to guide their learning. For instance, coded WCF can be confusing when students are not taught explicit grammar or statements like "counter arguments may be needed" may not be conducive to learning unless supported with classroom instruction.

Use consistent error codes: For WCF, I used a taxonomy of error categories adopted from Ferris (2006). Table 3.3 lists the error categories and error codes used in this study. The taxonomy was modified slightly to exclude the category of idioms due to the course curriculum discouraging the use of idioms in writing. I gave feedback on each essay following the guidelines above using the error codes in Table 3.3 and returned it to the students within one week. Because the AWE feedback focused on lower-level aspects of feedback, I only gave feedback on grammar for one body paragraph. Students were asked to use the provided feedback to revise and resubmit their essays.

TABLE 3.3
Error Categories and Codes

Error category	Error Code	Examples
Word choice	WC	Excluding spelling errors, preposition errors, pronouns, informal and unidiomatic usage
Verb tense	VT	Tense and aspect errors
Verb form	VF	Excluding verb tense errors
Word form	WF	Excluding verb form errors and verb tense errors
Articles	Art	The misuse of zero, definite, and indefinite articles
Singular-plural	#	Noun ending errors
Pronouns	Pro	The misuse of pronouns
Run-on	RO	Including comma splices
Fragment	Frag	Incomplete clauses
Punctuation	Punc	Inappropriate choice of punctuation marks. Excluded run-ons and fragments
Spelling	SP	Misspelled words
Sentence structure	SS	Including missing and unnecessary words and phrases and word order problems. Excluded run-ons and fragments
Informal	IN	Referring to register choices considered inappropriate for academic writing
Subject-verb agreement	SV	Excluding other singular-plural or verb form errors
Miscellaneous	?	Errors that could not be otherwise classified

Note: Error categories and codes. Adapted from "Does error feedback help student writers? New evidence on the short-and long-term effects of written error correction," by D. R. Ferris, 2006, p. 87.

In addition, in the current study, student-teacher conferences of ten to fifteen minutes were employed because previous literature has suggested teachers should facilitate machine feedback because students may not understand AWE feedback (Chen & Cheng, 2008; Li et al., 2015; Link et al., 2014; Wang et al., 2013). Oral conferencing has been a feature of many feedback studies (Cummins & Davison, 2007; Ferris, 1994; Hyland & Hyland, 2006; Lee, 2014a; Lee, 2014b) to encourage "active student participation and fostering learner autonomy" (Lee, 2017, p. 72) and is appropriate for the goals of the study.

3.4.3 Procedure

The comparison group submitted their essays to the teacher, where they received WCF and commentary as described above, followed by an oral conference. For the

experimental group, the students submitted their writing to AWE, where they had the option of revising their essays after receiving AWCF. When the students were satisfied with the overall result, they submitted the essays for the same WCF, commentary, and oral conference from the teacher, like the comparison group. Both groups could use the teacher feedback to improve their next essay. Table 3.4 shows the feedback the students received by group.

TABLE 3.4
Feedback Received by the Groups

	Comparison Group	**Experimental Group**
Writing Task 1	• Students submit the essay to the teacher. • Teacher provides WCF feedback using error codes for the introduction and the third body paragraph. Written commentary on content, organization, and development. • 15-minute oral conference	• Students submit their essay to AWE for feedback. • Students revise until they are satisfied with the results and submit to the teacher. • Teacher provides WCF feedback using error codes for the introduction and the third body paragraph. Written commentary on content, organization, and development. • 15-minute oral conference
Writing Task 2	• Students submit the essay to the teacher. • Teacher provides WCF feedback using error codes for the introduction and the third body paragraph. Written commentary on content, organization, and development. • 15-minute oral conference	• Students submit their essay to AWE for feedback. • Students revise until they are satisfied with the results and submit to the teacher. • Teacher provides WCF feedback using error codes for the introduction and the third body paragraph. Written commentary on content, organization, and development. • 15-minute oral conference
Writing Task 3	• Students submit the essay to the teacher. • Teacher provides WCF feedback using error codes for the introduction and the third body paragraph. Written commentary on content, organization, and development. • 15-minute oral conference	• Students submit their essay to AWE for feedback. • Students revise until they are satisfied with the results and submit to the teacher. • Teacher provides WCF feedback using error codes for the introduction and the third body paragraph. Written commentary on content, organization, and development. • 15-minute oral conference

3.5 Instruments

Four instruments were used to obtain the data for the study:

1. Being quasi-experimental, this study used five writing tasks as pre, post, and delayed

tests to investigate any changes in students' writing.

2. Three questionnaires were used to 1) collect demographic information, 2) examine changes in students" writing practices, and 3) investigate students' perception of using AWE in the classroom.
3. Two rating rubrics were used: an analytic rubric to rate the essays and a rating scale for analyzing the impact of hybrid feedback on students" drafts.
4. For the qualitative phase, a focus group interview was utilized to investigate students' engagement with and views of hybrid feedback.

3.5.1 Writing tasks

This study used pre, post, and delayed posttests to investigate any changes in students' writing quality after receiving hybrid feedback. Five writing samples were collected from each student from both experimental and comparison groups: pretest, in-class writing 1, in-class writing 2, in-class writing 3, and delayed posttest. The writing tasks used were single essay prompts selected from the Criterion online writing evaluation database. Criterion provides prompts of varying difficulties ranging from topics suitable for elementary 4th grade to graduate level. Because the participating students will be entering university, Criterion's 1st-year college level prompts were used in this study for the in-class writing and delayed posttests. For the pretest, a grade 12 high school prompt was used because this was the baseline for entering the program. These prompts aligned best with the three discourse modes addressed in the course: compare/contrast, cause/ effect and argumentative. For the pretest, an expository prompt was used because the participants were familiar with expository prompts. Table 3.5 lists the five writing prompts used in this study.

TABLE 3.5
Writing Prompts for 5 Writing Samples

Writing Test	Discourse Mode	Topic
Pretest	Expository	Successful students do well in school for many different reasons. Identify one or two important personal characteristics that help a student succeed in school. Use specific examples to show why you think these characteristics are important for student success.
In-class Writing Task 1	Cause and effect / Persuasive	The use of instant messaging, online social networks, e-mail and other forms of electronic communication has become increasingly common among people of all ages. How do these new technologies affect the way we socialize and build relationships? Explain your position with reasons and examples from your own experiences, observations or reading.
In-class Writing Task 2	Compare and contrast / Persuasive	In the ancient world, the term "liberal arts" referred to the education appropriate for free people (as opposed to slaves). In modern American higher education, the term is used to describe an education that focuses on general, rather than vocational, knowledge. Proponents believe that a liberal arts education is valuable because it prepares students for life by teaching them how to think. Opponents contend that the study of topics unrelated to one's professional path is a waste of time. Is a liberal arts education worthwhile? Develop your position by using evidence from your own experiences, observations or reading.
In-class Writing Task 3 (posttest)	Argumentative / Persuasive	Women's colleges, once common in the United States, have been going coeducational in increasing numbers in the past 40 years. Many people argue that women's colleges are unnecessary, now that all of the major United States colleges and universities are open to women. Others, citing studies that show that graduates of women's colleges are more successful than women who graduate from coed colleges, argue that women's colleges still have much to offer. Are single-sex colleges obsolete, or do they still provide an important alternative to coed colleges? Support your position with reasons and/or examples from your own experiences, observations or reading.
Delayed posttest	Argumentative / Persuasive	"Recently, major tobacco companies agreed to pay a financial settlement to several states, including California, for health problems caused by cigarette smoking and other kinds of tobacco addiction. This is unfair and unreasonable. Should car manufacturers be made to pay big settlements because people drive badly and have accidents? Should the makers of cell phones be held responsible for accidents people have because they are driving or using equipment while talking? No company should be made to pay because people misuse its products." —Edna Hacker Explain Hacker's argument and discuss the extent to which you agree or disagree with her analysis. Support your position, providing reasons and examples from your own experiences, observations or reading.

3.5.2 Student Questionnaires

Three questionnaires were used in the study. In the first questionnaire, the participants were asked about their demographic characteristics, including their IELTS and TOEFL

scores and years studying English. In addition, 12 questions asked students about their experiences with teacher feedback, their revision practices, and their familiarity with automated feedback systems with statements such as "I have previous experience with computer feedback systems (e.g., Grammarly, Microsoft word grammar, spelling checked, turnitin.com, etc.)" (See Appendix B for the Student Background and Perception Questionnaire). Responses are rated on a four-point Likert scale (Definitely agree = 4, Mostly agree = 3, Mostly disagree = 2, Definitely disagree = 1). The purpose of obtaining this information was to examine if students" previous feedback and revision practices affect hybrid feedback efficacy.

Second, a questionnaire about the students" cognitive processes when writing was administered in the beginning before the treatment and at the end of the program. Much literature in the field has shown that writing is not a linear process but rather involves multiple recursions of processes such as planning, translating, and reviewing (Flower & Hayes, 1981; Hayes & Flower, 1980; Hayes, 2012). Building upon the work of Shaw and Weir (2007) and an analysis of the key cognitive processes of L2 academic writers completing authentic writing tasks in a university context, Chan et al. (2017) created a questionnaire about L2 learners" writing processes (Chan et al., 2017; Chan, 2018). The questionnaire consists of 40 statements grouped under six cognitive phases of conceptualization, generating ideas, organizing ideas, generating texts, and monitoring. For example, statement 29 - "I usually check the accuracy and range of sentence structures and revise accordingly" is in the monitoring and revising at high-level phase. Responses are rated on a four-point Likert scale (Definitely agree = 4, Mostly agree = 3, Mostly disagree = 2, Definitely disagree = 1). This questionnaire was administered to the participants in this study at the beginning and end of the course (See Appendix C). See Table 3.6 for the questions and the cognitive phases they belong to in the questionnaire.

TABLE 3.6
Grouping of Questions for Cognitive Phases of Writing

Cognitive phases	Question item groupings
Conceptualisation	Q1, Q2, Q3, Q4, Q5, Q6, Q7, Q15, Q23, Q24
Generating ideas	Q8, Q9, Q10, Q11, Q22
Organizing ideas	Q12, Q13, Q14, Q19, Q21
Generating texts	Q16, Q17, Q18, Q20
Monitoring and revising at high-level	Q25, Q26, Q27, Q28, Q29, Q33, Q34, Q35, Q36, Q37
Monitoring and revising at low-level	Q30, Q31, Q32, Q38, Q39, Q40

Lastly, a Perception of Criterion Questionnaire was administered to students at the end of the course. The questionnaire, which is adapted from Dikli and Bleyle (2014, pp. 13-14), includes 14 Likert items regarding students' opinions about Criterion feedback, feedback on trait categories, and satisfaction with Criterion feedback. For example, "I found the Criterion feedback on Mechanics helpful (e.g., spelling, capitalization, punctuation)" elicits students' satisfaction with Criterion feedback on mechanics. Responses are rated on a five-point Likert scale (Strongly agree = 5, Agree = 4, Neutral = 3, Disagree = 2, Strongly disagree = 1). See Appendix D for a copy of the questionnaire on perceptions of Criterion.

3.5.3 Rubrics and Rating Scales

The EAP program where the study was conducted uses an adapted analytic version of the IELTS task 2 writing band descriptors (public version) to rate students' papers because, in the current curriculum, the students should be at IELTS band 6.5 by the time they finish the program. Therefore, the adapted version of the rubric focuses on Band 5, 6, and 7 of the IELTS writing task 2 band descriptors (See Appendix E for a copy of the program analytic rubric). The same analytic rubric was used in this study to rate the students' essays. The analytic rubric rates the essays in terms of task response, organization, lexical range and accuracy, and grammatical range and accuracy. Task response examines whether the topic is developed, and the essay addresses the prompt. Organization concerns the clarity and logical flow of ideas by using connectives and referencing.

Lexical range and accuracy examine the appropriacy of vocabulary and its effect on communication. Lastly, grammatical range and accuracy measure the sophistication of syntax and measure the impact of errors on communication. The rubric has five levels ranging from 1 to 5, with 1 indicating the essay was unintelligible or too short for assessment and 5 indicating the full realization of the Criterion.

To assess the impact of the AWCF and teacher feedback on students' revisions, a rating scale developed by Ferris (1997) to examine the significance and impact of revisions on writing was utilized (See Appendix F for the rating scale of revision). The scale was chosen because it considers both (a) the degree to which students used the feedback to revise their essays and (b) the impact of the revisions between the first and the lasat drafts on the overall quality of their essays on a six-point scale. For instance, the description of the highest score (6) reads, "substantive change(s) made by the student in response to comment, effect generally positive." The scale thus measures both the process and product of revisions.

Two raters/coders, who are highly experienced ESL teachers and with extensive experience instructing EAP writing classes, rated the writing responses collected for the study and coded the focus-group interview. The raters/coders participated in an orientation session before rating the essays and coding the focus-group interview, as described in Section 3.7.3.

3.5.4 Focus-group interviews

At the end of the course, a focus group interview was conducted with students in the experimental group to gather information and opinions from the students about their engagement with the teacher and automated feedback (see Appendix G for the student focus-group interview questions). A focus-group interview was chosen rather than an individual- interview because of organization issues: as soon as the course was done, many students went back to mainland China, and there was only limited time between the posttest and the end of class due to final exams. Although a face-to-face individual

interview may have higher potential for insights, the use of a focus group interview may have elicited more diverse opinions from the students by allowing for opportunities to state why students held a particular opinion or not by "piggybacking" on each other's responses (Patton, 1987, p. 135). Also, another advantage of focus groups compared to individual interviews is that they can also lead to a more natural and relaxed atmosphere for the participants than formal interviews (Marshall & Rossman, 2014). The focus group interview session was held with the experimental participants to elicit further their views on AWE feedback and lasted approximately an hour.

There were three broad themes for discussion during the focus group interview: students' perceptions of Criterion feedback, teacher feedback, and hybrid feedback. The questions were adapted from Li et al. (2015), which examined the impact of Criterion feedback on writing accuracy. For AWE feedback, students' views about the usefulness of trait feedback as well as the weaknesses and strengths of AWCF were elicited. For teacher feedback, students' views about the helpfulness of feedback and differences between AWE and teacher feedback were solicited. Lastly, for hybrid feedback, students' views about the effectiveness of hybrid feedback compared to teacher and AWE feedback alone were elicited. The focus group interview questions were structured to help reveal students' perceptions of using the hybrid approach to feedback as well as their beliefs, goals, and preferences, which could contribute to their behaviour as they work on their writing using automated corrective feedback.

3.6 Data Collection Procedures

Overall procedures for data collection for both the experimental and comparison groups can be found in the following timeline (Table 3.7). The writing tasks were given in weeks 1, 3, 5, and 7, and 3 months after the EAP class. The writing sample collected in week 1 was the pretest, the sample collected in week 7 was the posttest, and the sample collected 3 months after the end of the class was the delayed posttest. The paragraphs below detail the data collection procedures.

TABLE 3.7
Data Collection Timeline for Experimental and Comparison Group

Week	Comparison Group	Experimental Group
1	Student Demographic and Perception Questionnaire Student Writing Process Questionnaire Pretest	Student Demographic and Perception Questionnaire Student Writing Process Questionnaire Pretest Criterion Training Session
3	Writing Task 1	Writing Task 1
5	Writing Task 2	Writing Task 2
7	Writing Task 3 (posttest)	Writing Task 3 (posttest)
8	Student Writing Process Questionnaire	Student Writing Process Questionnaire Perception of Criterion Questionnaire Focus Group Interview
3 months after the end of EAP program	Delayed posttest	Delayed posttest

Before the study began, informed consent was received from all participants in the study following the Tri-Council ethics protocol (See Appendix H for informed consent forms and Appendix I for the ethics approval form). To ensure that the decisions to participate in the research are voluntary, a teacher not involved in the study was asked to meet with the students at the beginning of the semester and provide an informational session. This was intended to reduce the power differential if I was to recruit the participants.

During the information session, participant information sheets were distributed to brief the students/potential participants about the study and outline the project's scope and aims. Moreover, potential participants were assured that there were minimal risks in the study and were informed of the direct benefits of the study, including benefits for the participants as a result of exploring their own writing process, the indirect benefits for future courses using AWCF, and the advancement of the understanding of the integration of AWE feedback in the classroom. During this information session, the potential participants were given the opportunity to ask any questions to remediate

misunderstandings due to language proficiency. To give the potential participants sufficient time to consider the foreseeable risks and potential benefits, one week was given for students to decide whether to participate in the study or not.

The participants were told that they could withdraw from the study at any stage if they wished to do so and that their participation and withdrawal from the study would not affect their grades in the class to ensure that the potential participants were free of undue influence and/or coercion. To prevent bias, because I was in the dual role of researcher and teacher, all data was collected but not analyzed until after the final course grades were submitted at the end of the course. The participants were informed that the data would be destroyed for those participants who did not want to participate in the project. This was to ensure that there was no pressure on the students to participate in the study, and I would not know who was participating.

To ensure confidentiality and privacy, fictitious names and alphanumeric codes were used for the data. Also, any details of the participants which would make a participant easily identifiable were omitted or changed. The key to the codes was kept in a password-protected file system away from the data set to prevent unauthorized access. Participants were told that they would have the opportunity to see a summary of the results when requested. Finally, although the experimental group used hybrid feedback to augment teacher feedback, participants in the comparison group were not denied a benefit they already had since the teacher still gave them feedback; thus, the comparison group remained in their original state.

3.6.1 Writing Task Collection

For both groups, a pretest was administered in week 1, and three in-class writing tasks were collected in weeks 3, 5, and 7, with the third writing task in week 7 acting as the posttest. The experimental group submitted their writing tasks using Criterion. In contrast, the comparison group used Turnitin.com, an electronic writing submission program to check for plagiarism, as per program policy (Turnitin.com has a rudimentary

feedback system for grammar and spelling, but it was turned off). Because the experimental group would be submitting the writing tasks using Criterion, they underwent an in-class one-hour training on the use of Criterion in week one. The first 15 minutes were spent setting up student accounts and introducing the students to Criterion. The following 20 minutes were spent going over the tool's functionalities, and the last 30 minutes were spent uploading a sample writing task to explore the features and feedback. The students in the experimental group submitted the initial draft and up to 10 other drafts (the maximum number of drafts allowed by Criterion) to Criterion. The program kept a portfolio of holistic scores, trait scores, the number and type of errors from trait feedback, and the number of drafts and feedback for each student.

The students took the delayed posttest a semester later (three months after the EAP program ended). The three-month period was chosen because students would generally be at the end of a standard university semester to receive cumulative feedback from their courses and give course evaluations. Research has also shown that it is more likely that students may reflect on their overall learning during this time (Quinton & Smallbone, 2010). The students in the comparison group met together in the computer lab to write the final essay; however, for the experimental group, the timing of the delayed posttest coincided with the shuttering of the university due to the COVID-19 pandemic. During this time, many students left Canada for China due to the pandemic. Therefore, the delayed posttests for the experimental group were conducted through zoom over five weeks due to students' varying quarantine schedules and internet availability. The students were asked not to use any external sources while writing the delayed posttest essay and were reminded that there was no score attached to the essay.

3.6.2 Student Questionnaires

In week 1, all students in both groups were given the Student Background and Perception Questionnaire. The questionnaire collected data about the experimental group students' revision practices, previous experiences with feedback, and familiarity with AWE feedback systems. In weeks 1 and 8, the student Writing Process Questionnaire was given

to students in the experimental group. The questionnaire was administered before and after the treatment to examine changes in students' writing processes. Also, in week eight, on the last day of the class, the Perception of Criterion questionnaire was administered to the experimental group. Each of these questionnaires was completed in class using google forms.

3.6.3 Focus Group Interview Procedures

After completing the Perception of Criterion Questionnaire, students in the experimental group were asked to volunteer for the focus-group interview taking place the next day. The interview aimed to seek students' attitudes and perceptions concerning AWE and augment quantitative data gathered from the Writing Process Questionnaire and the Perception of Criterion Questionnaire. All students in the experimental group volunteered. The interview was informal, and my main role was to deliver guided questions for the group discussion and facilitate the smooth flow of their conversation by giving reminders and prompts for discussion.

3.7 Indicators of Writing Proficiency

To investigate the effects of hybrid feedback on the language, content, and organization of students" writing, I used a modified version of the analytic framework based on the Model of Writing Competence by Connor and Mbaye (2002), findings from previous research (Barkaoui & Hadidi, 2020), and measures used to evaluate writing in the EAP course. Grammatical, discourse, and strategic competencies were selected from the four competencies in the model because the fourth competence, sociolinguistic competence, is not explicitly taught in the course. This study examined the aspect of the pertinence of claims and argument quality, which is related to task response rating that assesses explanation of a concept with the relevant supporting ideas to explain the ideas fully. Syntactic and lexical range and sophistication were separated into two categories because they are taught explicitly as vocabulary and grammar in the course and match the marking rubric for writing in the course. Table 3.8 describes the corresponding

competence and constructs measured for human rating and computer analysis.

TABLE 3.8
List of Measures Used in the Study

Competence	Construct	Computer Analysis Measures	Human Rating
Grammatical Competence (Syntax)	Fluency	Number of words per essay	None
	Syntactic complexity	Global Syntactic Complexity Dependent Types of Noun Phrases (NP)	Rating on grammatical range and accuracy
	Linguistic Accuracy	Number of errors per 100 words	Rating on grammatical range and accuracy
Grammatical Competence (Lexis)	Lexical Complexity	Lexical Frequency Lexical Range Lexical Depth	Rating on lexical range and accuracy
Discourse Competence	Organization	Local Cohesion Global Cohesion Text Cohesion	Rating on organization
Strategic Competence	Pertinence of claims, warrants, and argument quality	None	Rating on Task Response

However, analytic ratings and trait scores may not be sensitive enough to measure writing improvement in one semester. Studies have shown that automated tools in writing can be used to investigate learner texts in regard to the linguistic properties and discourse components of the texts; they are faster and more consistent than manual human coding, thereby contributing to the validity and reliability of the results (Barkaoui & Hadidi, 2020; Crossley & McNamara, 2011; Crossley et al., 2011; Kyle & Crossley, 2017; Kyle & Crossley, 2018; Lu, 2011; Petchprasert, 2021). Therefore, this study used a combination of human rating and computer analysis to examine the quality of writing and its changes after receiving hybrid feedback in terms of both micro- and macro-aspects of writing. The human rating focused on evaluating the quality of specific aspects of writing. In contrast, the computer analysis focused on counting specific features, giving ratios, and corpus analysis (e.g., number of determiners, dependent clauses per T-unit, and age of acquisition of words, respectively). Previous research has shown that a combined approach effectively detects changes in writing features across proficiency

levels, tasks, learners, and time (Schiftner, 2013).

Computer tools that measure fine-grained indices were used to analyze several micro-features of the essays for both the comparison and the experimental groups, using various tools obtained from https://www.linguisticanalysistools.org. These include the Tool for the Automatic Analysis of Cohesion (TAACO) (Crossley et al., 2019b), the Tool for the Automatic Analysis of Lexical Sophistication (TAALES) (Kyle et al., 2018), and the Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASSC) (Kyle, 2016). These programs require using a specific corpus as a reference. Consequently, the academic subcorpora of Corpus of Con-temporary American English (COCA) was chosen because it is the largest corpora of Academic English (Davies, 2013).

For the selection of indices for various aspects of writing, first, the most common indices used in previous studies were identified through an extensive review of the literature; then, the indices were computed for each essay, and a correlation table was created to examine if any indices were highly correlated. Indices with high correlations were removed. If there were no high correlations, the frequency of occurrence of indices was examined, and the index was removed if most of its values were close to zero. The following paragraphs provide more detail on the various measures and indices used in the study. The analysis of grammatical competence by computer analysis and human rating contributes to answering questions about changes in language in students' writing; the analysis of discourse and strategic competence answer questions about changes in organization and content, respectively (see Appendix J for a full list and descriptions of automated indices).

3.7.1 Grammatical Competence

Grammatical competence was examined in relation to fluency, linguistic accuracy, lexical complexity, and syntactic complexity.

3.7.1.1 Fluency

Fluency is essential for measuring language development in writing because, in the literature, increased fluency indicates more cognitive resources for higher-level processing resulting in higher-level content (Deane & Zhang, 2015). Previous research shows that the number of words in writing is the strongest predictor of essay quality, and higher proficiency in writing correlates with longer texts (Crossley & McNamara, 2016). Although previous research operationalized fluency as the average number of words per T-unit (e.g., Cumming et al., 2006; Wolfe-Quintero et al., 1998), the average number of words per T-unit may be a measure of syntactic complexity and not fluency (Norris & Ortega, 2009); thus, in this study, the number of words per essay was used to operationalize fluency.

3.7.1.2 Linguistic Accuracy

Studies on linguistic accuracy have found a significant positive relationship between linguistic accuracy and writing performance (Wolfe-Quintero et al., 1998). Although there are many different methods of measuring the number of linguistic errors, such as error-free T-unit ratio and error-free clauses per all clauses in the text, they are labour intensive to compute, and human raters may have difficulty reliably identifying, classifying, and evaluating linguistic errors (Cumming et al., 2006). Although studies of precision and recall of AWE systems" error detection have found precision to be high, meaning that the errors these systems locate are high; while recall was low, meaning that they fail to detect some errors (see Burstein et al., 2004), research has shown that most errors that AWE systems miss were punctuation errors, and the analysis of linguistic accuracy by AWE systems has a strong association with human ratings of overall grammar and mechanics errors (Crossley et al., 2019a). Therefore, in the current study, Criterion was used to identify linguistic errors. Criterion identifies four types of errors: grammar, usage, mechanics, and style. Errors of grammar, usage, and mechanics identified by Criterion were used to measure linguistic accuracy in this study. The frequency of errors was calculated as errors per 100 words.

3.7.1.3 Lexical Complexity

Lexical complexity is an important aspect of academic writing (Storch, 2009), and it is multidimensional (Bulté & Housen, 2014; Read, 2000). Previous research has argued for measuring different dimensions of lexical complexity such as lexical richness, diversity, and sophistication (Lu, 2012; Read, 2000).

Lexical Richness (Lexical Frequency): Most lexical frequency indices depend on frequency lists and are based on the hypothesis that a higher lexical proficiency results in the use of less frequent words (Meara & Bell, 2001). Word frequency has traditionally been assigned to the breadth of knowledge category, but this categorization is debatable. Ellis (2002), for instance, argues that the production and comprehension of words is a function of their frequency of occurrence in language. Within this perspective, word frequency affects lexical acquisition because a word's repetition strengthens the connection between the word and its meaning categorization. As learners are exposed to frequent words, there is a reduction in processing time because the practice time with the word increases. Such a model of the acquisition of lexical richness or lexical frequency is supported by studies that demonstrate that high-frequency words are named and processed more quickly than low-frequency words (Balota & Chumbley, 1984; Kirsner, 1994).

Studies on L2 writing proficiency found that L2 learners with lower proficiencies are more likely to comprehend, process, and produce higher frequency words (Crossley & Salsbury, 2010; Ellis, 2002) compared to advanced learners (Meara & Bell, 2001). Vidakovic and Barker (2010) found that both the frequency and diversity of lexical bundles increase with proficiency. Lexical bundles are combinations of a number of words that frequently occur together in discourse. An example of a bigram is "number of," and of a trigram would be "turn the page." Vidakovic and Barker found that lexical bundles are rarely used by lower proficiency learners, but for intermediate and advanced learners, the bundles become more varied and frequent. The following are common indices for lexical richness (lexical frequency): Frequency All Words (AW), Frequency Content Words (CW), Frequency Function Words (FW), Bigram Frequency, and Trigram

Frequency. However, the occurrence of Trigrams in the data set was very low; therefore, it was removed. Also, the frequency of all words and function words were highly correlated (r = .92), resulting in Frequency FW being removed.

Lexical Diversity (Range): The premise behind lexical diversity indices is that a more diverse vocabulary indicates a more proficient and more extensive lexicon. Historically, indices that measure lexical diversity concentrated on type-token ratios (TTR), which divides the number of different words (types) by the total number of words (tokens) in a text. However, a common problem with TTR is that texts with a higher number of tokens give lower values of TTR because the writer uses more function words (Johansson, 2009). Therefore, a better method is to determine the reference corpus frequency of each word that occurs in a target text and then create an average frequency score for a text by dividing the sum of all word frequency scores by the total number of words in a text (Kyle & Crossley, 2015). Kyle and Crossley (2016) found that lexical range, bigram, and trigram features were predictive of independent writing quality, and lexical range is typically calculated for AW, CW, and FW. The following are the standard indices for lexical diversity (lexical range) in the literature: Range AW, Range CW, Range FW, Bigram Range, and Trigram Range. However, the occurrence of Trigrams in the texts in this study was very low; therefore, it was removed. The range of AW and CW was also highly correlated (r = .83), resulting in the range for CW being removed.

Lexical Sophistication (Depth): The sophistication of vocabulary has been measured in the literature by hypernymy, word polysemy, age of acquisition index, N-gram strength of association, and in academic settings, by the academic word list. The following paragraphs describe, in-depth, the indices for lexical sophistication.

Word Hypenymy: Hypernymy shows the relationship between a generic term (hypernym) and a specific instance of that term (hyponym). For example, a hypernym of apple would be fruit, and the hyponym of fruit would be an apple. From an L2 acquisition perspective, hypernymic relations are acquired as the learners acquire more specific lexical

knowledge. The overuse of hypernym would result in inappropriate over-generalizations in writing (Wolter, 2001). A low hypernymy value for a text reflects an overall use of less specific words. In contrast, a higher hypernymy value reflects an overall use of more specific words (Crossley et al., 2011). Hypernymy is usually estimated for nouns and verbs separately (Kyle et al., 2018). These two indices were not highly correlated (r = -.06) in the current dataset; therefore, both were kept.

Word polysemy: Polysemous words are words that have multiple meanings. For instance, the word "study" has at least eight related senses or meanings, including the devotion of time and attention to gaining knowledge, a detailed investigation, a course of study, a room used for doing academic work, a piece of work done as an experiment, etc. (Verspoor & Lowie, 2003). From an L2 acquisition perspective, word polysemy can affect how efficiently L2 learners recognize meaning relationships between a word's senses (Verspoor & Lowie, 2003). Studies concerning the polysemy knowledge of L2 learners have found that word sense knowledge increases as L2 learners gain proficiency (Crossley & Salsbury, 2010). Word polysemy indices are usually broken down to adjectives, adverbs, CW, nouns, and verbs in the literature (Kyle et al., 2018). However, polysemy and hypernymy for nouns and verbs were highly correlated (r = 1), and verb polysemy was removed.

N-gram strength of association: A critical component of writing ability is using words appropriately in context. Strength-of-association norms measure the conditional probability that words will occur together (Kyle et al., 2018). For example, bigrams such as "optimistic about" are more strongly related than the ones in "and the" and "in the." From an L2 acquisition perspective, the N-gram strength of association is associated with collocational knowledge and is an essential aspect of lexical proficiency in L2 contexts (Bestgen & Granger, 2014). Staples et al. (2013) analyzed lexical bundles in TOEFL iBT writing section and found that the highest scoring responses contained less repetitive lexical bundles. N-gram strength of association is computed for bigrams and trigrams, with trigrams being computed in two ways. The first way is when the first word is

considered item 1 and the following bigram is considered Item 2, or the second way, where the first bigram is considered item 1 and the third word is considered item 2. However, the three indices were highly correlated ($r > .83$), so only Bigram Association Strength was kept.

Academic Wordlist (AWL): Research has shown that more proficient writers use more academic words when responding to academic writing tasks. Higher AWL values indicate more sophisticated vocabulary use (Laufer & Nation, 1995). AWL value is calculated by counting the number of AWL words in the text divided by the number of words in the text.

Contextual Distinctiveness: Contextual distinctiveness measures the diversity of contexts in which a word is encountered (Kyle et al., 2018). It is operationalized as co-occurrence counts of words that are located in the immediate environment. For example, the word "amok" is constrained by its immediate environment (usually only occurring in the bigram run amok); however, run has fewer constraints than amok and occurs with a variety of different words (McDonald & Shillcock, 2001). Crossley et al. (2013) found contextual distinctiveness to be a more powerful predictor of lexical development and theoretically more compelling than word frequency measures because it is based on a word's lexical environment. The selected indices use latent semantic analysis to measure the semantic association between words through a mathematical technique that calculates associations between words and the context in which they occur (Crossley et al., 2013). It is calculated by averaging Latent Semantic Analysis (LSA) cosine scores for all related words for each word in the text. LSA is a mathematical and statistical technique based on large corpora of texts to measure semantic similarity in meaning. Therefore, the automated index of contextual distinctiveness was chosen.

In summary, this study follows the analysis of lexical features that past studies have determined in terms of richness (depth), diversity (range), sophistication (depth) and distinctiveness are important indicators of L2 lexical growth and proficiency (Crossley et

al., 2014; Kyle & Crossley, 2016; McNamara, Graesser, McCarthy, & Cai, 2014). The following are the selected indices for lexical sophistication: AWL value, Bigram Association Strength, Hypernymy Nouns, Hypernymy Verbs, Polysemy Adjectives, Polysemy Adverbs, Polysemy CW, and Contextual distinctiveness. TAALES was used to estimate the above indices.

3.7.1.4 Syntactic complexity

A key measure of L2 proficiency has been syntactic complexity (Bulté & Housen, 2014; Lu, 2011). Like lexical complexity, syntactic complexity is also multidimensional (Lu, 2011; Norris & Ortega, 2009). Syntactic complexity refers to the levels of sophistication or variety/range of structural forms in language production (Ortega, 2003). From an L2 acquisition perspective, as learners become more proficient, they produce longer and more varied and sophisticated syntactic structures (Norris & Ortega, 2009). While the literature agrees on the general definition of syntactic complexity, there is some disagreement on measuring it (Biber et al., 2020). Lu (2011) divided measures of syntactic development into five groups of specific indices: 1) length of production, 2) sentence complexity, 3) subordination, 4) coordination, and 5) particular structures. However, these global measures have been criticized because they are not very interpretable, confound different syntactic categories, or measure similar things (Guo et al., 2013; Qin & Uccelli, 2016).

In a study by Kyle and Crossley (2018), the authors found that phrasal complexity indices are better indicators of writing quality than both global and clausal indices because some critics argue that these measures do not reflect syntactic and semantic complexity accurately enough (Ishikawa, 1995). In studies that examined academic texts, in particular, it was found that complex noun phrase (NP) constituents and complex phrases are more frequent than other types of structures (Biber et al., 2011). Corpus-based research suggests that phrasal (not clausal) complexity is a crucial distinguishing feature of academic writing, especially for first-year university undergraduate students, which is similar to the target of the current teaching context (see Staples et al., 2016). Thus, in

current research, the global complexity measures from Lu (2011) were used in addition to measures of dependent types of noun phrases. Global complexity measures are based on ratios with sentences, clauses, or t-units as the denominator, but in the current study, T-units were chosen as denominators because Yang et al. (2015) found T-unit complexity to be significantly and positively correlated with writing quality scores compared to clauses or sentences. Another reason for using T-units for selected indices was because T-units and Sentences had a high correlation in the current study. For example, Mean length of sentences (MLS) and MLT have a correlation of .79, and dependent clauses per clause (DC/C) and DC/T had a correlation of .95. There are five global complexity measures: Length of production unit - mean length of T-units (MLT), Sentence complexity - sentence complexity ratio (C/S), Amount of subordination - dependent clauses per T-unit (DC/T), Amount of coordination - coordinate phrases per T-unit (CP/T), and Degree of phrasal sophistication - complex nominals per T-unit (CN/T). None of them were highly correlated, so the following indices were chosen: MLT, C/S, DC/T, CP/T, and CN/T.

For measures of dependent types of noun phrases, determiners, adjectival modifiers, prepositional phrases, possessives, verbal modifiers, nouns as modifiers, relative clause modifiers, adverbial modifiers, use of "and" as a conjunction, and use of "or" as conjunction were identified as important (Kyle & Crossley, 2018). A correlation analysis was performed to reduce the number of indices, but the analysis did not show multicollinearity. However, descriptive statistics indicated that six of the indices in the dataset were close to zero and were removed (verbal, nouns, relative and adverbial modifiers; "and" and "or" as conjunction). Therefore, only *determiners, adjectival modifiers, prepositional phrases,* and *possessives* were used for analysis. TAASSC was used to estimate the above indices.

3.7.2 Discourse Competence

Halliday and Hasan (2014) were among the first to note that writers use cohesive devices to create a discourse between the writer and the reader. Moreover, Scott (1996) found that many L2 learners have difficulty understanding how cohesive and logical ties are

constructed in a text and that L2 instruction needs to address this explicitly. Much literature has shown that expert judgements of essay quality are correlated with how the text is organized, and cohesion is one of the strongest predictors of essay quality (McNamara et al., 2010). Although cohesion and coherence may be synonymous in many rubrics and the mind of a human rater, the distinction between the two is important for automated evaluation. Cohesion refers to the presence or absence of explicit cues in the text, while coherence refers to the understanding the reader derives from the text. Cohesion can be measured based on linguistic features found in the text, while coherence of the text may interact with the readers' background knowledge (O"reilly & McNamara, 2007). Crossley et al. (2016) explained that the explicit cues that allow the readers to make connections between ideas can be global (between paragraphs), local (between sentences), and text (throughout the entire text).

In addition, global, local and text cohesion can be further divided into coreference and conceptual cohesion. Coreference cohesion occurs when a noun, pronoun, or noun phrase refers to another constituent in the text (McNamara et al., 2010). Crossley and McNamara (2011) found that many global cohesion indices were significantly and positively correlated with essay quality scores. For example, noun overlap for global cohesion measures the proportion of paragraphs in a text where there are overlapping nouns. In contrast, conceptual cohesion concerns the extent to which the content of sentences or paragraphs is similar semantically or conceptually. The main measures of conceptual cohesion are based on latent semantic analysis (LSA). LSA is a mathematical and statistical technique based on large corpora of texts to measure semantic similarity in meaning or conceptual relatedness between words, sentences, and paragraphs that may not be related morphologically (Crossley & McNamara, 2012).

There are many indices that can be used for coreference global cohesion. For instance, the overlap between all words, nouns, verbs, and arguments across adjacent or two paragraphs are indices of coreference global cohesion. Correlation between indices was examined to reduce the number of indices, and adjacent paragraph overlap for all words

was selected because it was highly correlated with other indices (r= .79 with adjacent paragraph overlap of nouns, .83 with adjacent paragraph overlap of arguments). Likewise, the overlaps between adjacent and two paragraphs were all highly correlated, so they were removed. However, for conceptual cohesion, there was no correlation between noun and verb overlap between paragraphs. Thus, both were kept. The following are the selected indices for global cohesion: adjacent paragraph overlap for all words, conceptual overlap of verbs between paragraphs, and conceptual overlap of nouns between paragraphs.

While global cohesion is wildly recognized as being indicative of proficient writing, findings for local cohesion measures are not as clear-cut. L1 research of English has shown that less proficient writers use more explicit cohesion devices to link sentences together (McCutchen & Perfetti, 1982). Likewise, a recent study by Crossley and McNamara (2012) and Guo et al. (2013) confirmed that the production of local cohesion devices (e.g., content word overlap and conditional connectives) in L2 writing was also negatively correlated with essay quality ratings. However, some researchers have argued that in many ESL writing classrooms including EAP classes, students are explicitly taught and encouraged to use local cohesive devices and sentence transitions; these cohesive devices would be more valued and tend to correlate with higher writing scores (Chiang, 2003; Liu & Braine, 2005). Therefore, local text cohesion is included in the study. Although the global and local indices are similar (between paragraphs and between sentences, respectively), the correlation pattern differed. Adjacent sentence overlap for all words was correlated less than .70 for noun, verb, and argument overlaps. In addition, adjacent sentence overlap for nouns was correlated (.85) with adjacent sentence overlap for all arguments, so only the index for noun overlap was kept. However, correlations between all words, verbs, nouns, and arguments between adjacent sentences and adjacent two sentences were high, so only the indices between adjacent sentences were kept. Lastly, the coreference index of adjacent sentence overlap for all words was highly correlated with the syntactic overlap of nouns between sentences (.71), so it was removed. The following are the indices for local cohesion included in this study: adjacent

sentence overlap for all words, adjacent sentence overlap for nouns, adjacent sentence overlap for verbs, and conceptual overlap of verbs between sentences.

Text cohesion concerns cohesion throughout the text. Examples of text cohesion include the use of connectives throughout a text and givenness of information. Givenness concerns the given information that has been presented earlier in the text, and processing given information can be more accessible because it was already mentioned (McNamara et al., 2013). In other words, givenness reflects "the amount of information that is recoverable or repeated from the preceding discourse" (Crossley et al., 2015b, p. 3). Givenness indices based on the number of content words that are repeated and on pronoun density calculated at the text level have been found to be a positive indicator of text coherence in previous studies (Crossley & McNamara, 2011; Crossley et al., 2015a; Crossley et al., 2019b). Lastly, research has shown that connectives provide important information about a text's cohesion (Crossley et al., 2016; Halliday & Hasan, 2014). The three indices were not shown to have multicollinearity for the current data set. The following are the selected indices for text cohesion: Repeated content words, Pronoun density, and All connectives. TAACO was used to estimate the above indices.

In summary, Table 3.9 lists the selected indices of grammatical and discourse competence used in this study.

TABLE 3.9
Selected Indices for Automated Analysis

Measures	**Selected indices**
Fluency	Number of words per essay
Global Syntactic Complexity	MLT C/S DC/T CP/T CN/T
Dependent Types of Noun Phrases (NP)	Determiners Adjectival Modifiers Prepositional Phrases Possessives
Number of errors	Number of errors per 100 words
Lexical Frequency	Frequency AW Frequency CW Bigram Frequency
Lexical Range	Range AW Range FW Bigram Range
Lexical Depth	AWL Value Bigram Association Strength Hypernymy Nouns Hypernymy Verbs Polysemy Adjectives Polysemy Adverbs Polysemy CW Contextual Distinctiveness
Local Cohesion	Adjacent Sentence Overlap for All Words Adjacent Sentence Overlap for Nouns Adjacent Sentence Overlap for Verbs Conceptual Overlap of Verbs Between Sentences
Global Cohesion	Adjacent Paragraph Overlap for All Words Conceptual Overlap of Verbs Between Paragraphs Conceptual Overlap of Nouns Between Paragraphs
Text Cohesion	Repeated Content Words Pronoun Density All Connectives

3.7.3 Human Rating

Two raters who had 11 and 17 years of experience teaching EAP and writing rated each essay according to task response, organization, lexical range and accuracy, and grammatical range and accuracy using the EAP analytic rubric (see Appendix A).

Although both raters had previously used the rating scale, they both went through a formal standardization session to clarify the rating criteria, establish anchor papers, and provide standardization (Davidson, 1991; Erdosy, 2004). Prior to rating the papers, the raters reviewed the writing tasks and the rubric. After receiving instruction on the rubric, the raters practiced using the rubric on a sample set from a previous cohort written on the same prompts as the essays in the current study. In the training session, the raters discussed the rubrics and rated the training set to become familiar with the rating scales and clarify any disagreements. During this discussion, raters discussed the criteria and score levels on the rating scale and any unclear points. The raters discussed the writing prompt concerning task achievement to see what a fully developed position might require in relation to the prompt. Next, the raters individually scored sample responses, and any discrepancies in ratings and interpretations of the rubric were discussed and resolved. Studies have shown that resolving score differences in the rating of writing samples by discussion improves the accuracy of scores (Johnson et al., 2005). The norming session continued until the raters' adjacent agreement was 70% in a set of 10 writing samples, which was needed for ratings to be considered reliable (Stemler, 2004).

Before essays were assigned to raters, the experimental and comparison group's writing samples were randomly assigned to different sets with the students' names redacted, so the students' group designation did not influence the raters. Also, the writing samples before and after feedback were randomized into different sets for the experimental group to decrease rater bias. In addition, before rating revisions, the raters underwent a similar norming session. To reduce fatigue, each rater marked essays for up to 30 minutes at a time only. Each Criterion was rated separately to avoid raters becoming familiar with the essays and thus creating a halo effect. After rating each Criterion, the raters discussed and resolved discrepancies.

Inter-rater agreement was computed for each Criterion in terms of the percentage of exact agreement and Cohen's Kappa. Only one rating had a discrepancy of more than 0.5, and the adjacent agreement, the percentage of times two or more raters give a score within

one level of each other on ratings of performance, was at 100 percent for all ratings. As Table 3.10 shows, the exact inter-rater agreement was above 75% for all four criteria.

TABLE 3.10
Inter-rater Reliability for Human Ratings

Criterion	% Exact Agreement	Kappa
Task Response	78.60	0.72
Organization: Coherence and Cohesion	89.30	0.84
Lexical Range and accuracy	87.44	0.82
Grammatical Range and accuracy	82.32	0.75

3.8 Data Analysis Procedures

3.8.1 Dataset

The writing samples for the study came from two cohorts: the experimental and comparison group. Both groups wrote five essays at different time points (weeks 1, 3, and 5, 7, and three months after the EAP program), with the first essay being the pretest and the last essay being the delayed posttest. There were 80 essays from the comparison group and 84 from the experimental group for new pieces of writing for a total of 164 writing samples. One participant from the experimental group was unable to write the delayed posttest. In the experimental group, for writing tasks 1, 2 and 3, the first and last drafts of the tasks were collected for a total of 102 writing samples. In total, there were 216 writing samples between the two groups. Data also included the responses of the experimental group students to the writing processes questionnaire at the beginning and at the end of the courses; to the questionnaire about their perceptions of hybrid feedback; and to the focus group interview.

3.8.2 Data Cleaning

All writing tasks from both groups were manually converted to text files so that the automated writing analysis tools could process them. However, due to differences in operating systems, operating language, and Unicode (encoding of characters) used, each

file had to be stripped of all extraneous encoding. For example, some students used tab to indent while others used spaces, some used line return while others used new line or carriage return for new paragraphs, and some characters were encoded in Chinese characters. Two software programs were used to clean and strip the text and reformat them in Unicode: TextSoap and CleanText. In addition, all titles and headings were removed along with the writing prompts from the text files to have the most accurate word count for data analysis.

3.8.3 Data Analysis Procedures

The following sub-sections describe the qualitative and quantitative data analysis procedures for each research question. Table 3.11 shows the data type, data source, and data analysis for each of the research questions of the study.

TABLE 3.11

Data Analysis procedures for the Research Questions.

Research Questions	**Data Type**	**Data source**	**Data Analysis**
1. How does the use of hybrid corrective feedback affect the students" approaches to writing compared to students who only receive teacher feedback?	Qual	Focus-group interview	Thematic analysis
	Quan	Rating scale for Changes in Revision from (Ferris, 1997, p. 322) Writing Process Questionnaire pre and post treatment	Descriptive and inferential statistics Repeated measures MANCOVA
2. How does the use of hybrid corrective feedback affect the language, content, and organization of students" writing compared to students who only receive teacher feedback?	Qual	Focus-group interview	Thematic analysis
	Quan	Automated measures and human ratings for grammatical competence Automated measures and human ratings for discourse competence and human ratings for strategic competence	Descriptive and inferential statistics Mixed MANOVA or ANOVA for between groups and repeated measures MANCOVA or ANCOVA for within groups
3. How do the students view hybrid AWE corrective feedback?	Qual	Focus-group interview	Thematic analysis
	Quan	Student perception Questionnaire	Descriptive statistics

To address RQ 1 (How does the use of hybrid corrective feedback affect the students" approaches to writing compared to students who only receive teacher feedback?), responses to the Writing Process Questionnaire and ratings of revisions were analyzed as

follows.

This study operationalized students' writing approach in terms of a questionnaire on the five cognitive processes of academic writing (Chan et al., 2017): conceptualization, generating ideas, organizing ideas, generating texts, and monitoring and revising texts. The questionnaire was administered to students in the treatment group before and after the treatment to explore changes in their writing practices. To analyze the questionnaire data, a mixed MANOVA comparing changes in perceptions of writing processes between the comparison and experimental groups was used. The dependent variables were the individual indices for each cognitive phase computed as a scale, and the between subject variable was the group. The within-subject factor was the two time periods. In addition, descriptive statistics were calculated for the rating scale for significance and impact of revisions. The rating scale examined the impact of the revisions between the first and the last drafts on the overall quality of writing. Reliability analysis was carried out on each phase of the writing process for pretests and posttests. Literature indicates that the generally accepted rule is that 0.6-0.7 indicates an acceptable level of reliability, and 0.8 or greater is a very good level (Ursachi et al., 2015). As seen in Table 4.33, alpha for all scales was above .60.

To answer RQ2 (How does the use of hybrid corrective feedback affect the language, content, and organization of students" writing compared to students who only receive teacher feedback?), two primary statistical analyses were carried out: (a) comparisons of grammatical and discourse indices for new pieces of writing between comparison and experimental groups and (b) comparisons of grammatical and discourse indices between the first and the last drafts for writing tasks 1, 2, and 3 for the experimental group. The two analyses were conducted to account for criticism of feedback not examining the longitudinal changes on new writing pieces to reflect retention of the feedback (Truscott, 1996).

To analyze changes in new pieces of writing between groups, a mixed MANOVA

between the experimental and comparison groups was used to see if there were any significant differences between the experimental and comparison groups. The dependent variables were the individual grammatical and discourse indices, and the between subject variable was the group. The within-subject factor was time with five levels (Week 1, 3, 5, and 7, and 3 months after the EAP program). Because one student did not write the delayed posttest, the student was removed from between group analysis. This resulted in the selection of Pillai"s Trace test for MANOVA because the removal resulted in the sample sizes being equal, and the Pillai"s Trace has the most power, the least error and the greatest robustness to violations of test assumptions when the samples sizes are equal and small (Field, 2018).

The data was checked to meet the six assumptions that are required for a mixed MANOVA: 1) the dependent variables are continuous variables, 2) the independent variable is categorical (group), and the participants were measured at all time points, 3) the number of participants is greater than the number of dependent variables, 4) there were no univariate or multivariate outliers, 5) there is a linear relationship between each pair of dependent variables for each related group of the independent variable, and lastly, 6) There is no multicollinearity, I.e., no correlation higher than .90. As detailed in section 3.7, all indices with correlations higher than .80 were removed. A significant MANOVA result is traditionally followed by a separate ANOVA analysis of each of the outcome variables; the overall multivariate test protects against inflated Type I error rates because if the multivariate test is non-significant, then any subsequent tests are ignored. However, Harrison et al. (2020) write that "this notion of "protection" is a little misleading because a significant MANOVA often reflects a significant difference for one rather than all dependent variables" (p. 341). Therefore, to protect against Type 1 error, the p-value was corrected by applying the Bonferroni correction by dividing the p-value by the number of dependent variables (Field, 2018). The follow-up analysis of mixed ANOVA was conducted on each dependent variable separately, adjusting the p-value by using Bonferroni correction.

For comparisons between the first and the last drafts, a repeated measures 2x3 MANCOVA was used. The dependent variables were individual grammatical and discourse indices, and the independent variable was draft, with 2 levels, first and last. The within-subject factor was the three-time periods. The number of drafts submitted was used as a covariate. However, the analysis found no main or interaction effects for the covariate. Therefore, the analysis was done again without the covariate to simplify the interpretation.

For both, comparison between groups at five-time points for both groups and comparison between the first and the last drafts at three-time points for the experimental group, if the analysis examined only one index, either a mixed ANOVA or a repeated measures 2x3 ANOVA was used. For these analyses, Mauchly"s test statistic was used to ensure that the variances of differences were not significant, which is needed to validate a repeated measures analysis of variance. If Mauchly"s test statistic was significant, Greenhouse–Geisser correction was used to adjust for lack of sphericity (Field, 2018).

To answer RQ 3 (How do the students view hybrid AWE corrective feedback?), descriptive statistics were used to analyze the students' answers to the Student Background and Perception Questionnaire. In addition, the focus group interview was transcribed verbatim and inductively coded utilizing grounded theory to identify patterns in the data following procedures outlined by Lincoln and Guba (1985). The overriding patterns were identified and grouped, and then the data was divided into smaller and more meaningful units using Nvivo 12.1. I met with an independent researcher during the iterative process of refining the coding scheme to identify emerging themes and sub-categories to provide a layer of peer debriefing and improve the trustworthiness of the qualitative analysis (Denzin & Lincoln, 1994).

Data were coded using NVivo 12, and then thematic analysis was used to interpret the qualitative data. Initial themes were identified from the quantitative findings. To achieve validity and reliability in interpreting the data, two peer researchers coded the interview

data independently. The data analysis was conducted using a modified version of six-phase process for thematic analysis recommended by Braun and Clarke (2006). The data collection procedure for the interview followed the first four steps, while the results sections discuss the last two. The following paragraphs detail the first four phases:

1. Familiarizing yourself with your data:
The interview was first transcribed by using Otter (https://otter.ai), an AI-powered transcription software. I verified the resulting transcript, and corrections were made to any errors by comparing it to the original audio recording. To increase reliability, another researcher repeated the verification process by reading the transcript and checking the transcription against the audio file for accuracy. The process of transcription and the verification helped to familiarize the researchers with the data.

2. Generating initial codes:
Both researchers independently coded the interviews by tagging selections of the transcript in Nvivo 12. To preserve the context of the code, extracts of data were highlighted, and each section was coded multiple times as needed. Codes were produced based on the semantic content of the interview. For example, After the initial coding was done, the researchers discussed each code systematically and changed categories and coding as necessary.

3. Searching for themes:
Both researchers independently combined the codes to form overarching themes and accounted for unexpectedly salient or emergent themes. A thematic map of perceptions of hybrid feedback and its effect on the writing process was developed by utilizing MindNode 7.0.3.

4. Reviewing themes:
Collaboratively, coded data extracts were reviewed systematically by collating data extracts for each theme for best coherence. Next, the validity of individual themes in

relation to the data set was explored by examining if the thematic map reflects the data set as a whole.

3.8.4 Selected Individual Case Analyses

Previous research on teacher and automated feedback found that different students engage with AWE feedback differently due to individual factors (e.g., Zhang & Hyland, 2018; Zhang & Hyland, 2018). The individual case analyses may provide more insights into students' engagement and perception of hybrid feedback and may help to explain and elaborate on some possible causal links between perception, engagement, and utilization of feedback. To investigate students' perceptions, more than one case is needed to compare two varying attitudes to hybrid feedback

By combining multiple data sources to build a more complete picture of the three selected cases, these qualitative analyses aim to offer a more nuanced understanding, compared to the holistic group results, of individual differences in students' perception, engagement, and utilization of feedback. In addition to the analysis procedures in the main study, the writing samples for the selected students were examined at greater depth for more insights in relation to research questions 2 and 3: why students writing process changed and reasons for their perception of hybrid feedback.

By examining the results of the Perception of Criterion Questionnaire, the focus group interview, and the ratings on the Significance and Impact of Revisions, three cases that would be representative of high preference for and engagement with teacher and machine feedback, preference for teacher feedback only, and preference for machine feedback were selected.

In the next chapter, the results of these different analyses are reported.

Chapter 4: Results

This chapter presents the results of the analyses as they pertain to the four research questions of the study.

4.1 Changes in Students' Approaches to Writing

Change in students' approaches to writing was examined by comparing students' responses to the Writing Process Questionnaire before and after the course and a second questionnaire about students' background and perceptions of feedback in previous classes.

The student Background and Perception of Feedback Questionnaire showed that the students in both groups had similar experiences with teacher feedback and computer feedback systems. The only statement on which the two groups were different by more than 0.5 was, "I found peer feedback helpful in revising my essays" for the comparison group. However, peer feedback was not part of the focus of this study. See Table 4.1 for the descriptive statistics for students' prior experience with feedback by group.

TABLE 4.1

Descriptive Statistics for Students" Prior Experience with Feedback

	Experimental		Comparison	
	M	**SD**	**M**	**SD**
I think doing more writing is important to improve my writing.	3.24	0.66	3.31	0.60
I pay attention to the score when my writing is returned.	3.29	0.77	3.63	0.62
I pay attention to the feedback when my writing is returned.	3.35	0.61	3.88	0.34
I think the feedback I received from my instructors was timely.	3.41	0.71	3.81	0.40
I try to avoid similar problems in future writing when I receive feedback.	3.29	0.59	3.81	0.40
I revise my essays before submission.	3.12	0.70	2.94	0.85
I think revising my essays is an important part of the writing process.	3.35	0.61	3.56	0.63
I like revising my essays.	2.88	0.93	2.69	1.01
I find instructor feedback helpful when revising my essays.	3.53	0.51	3.69	0.48
I find peer feedback helpful in revising my essays.	2.76	0.66	3.31	0.70
I have previous experience with computer feedback systems (e.g. Grammarly, Microsoft word grammar and spell checker, turnitin.com, etc.…)	2.82	0.81	2.69	0.93
If yes to the previous question, I find computer feedback helpful in revising my essays.	3.06	0.75	3.25	0.77

(N=16 for comparison, 17 for experimental)

The Writing Process Questionnaires were conducted for the two groups at two time points: in week 1, prior to treatment and in week 8, at the end of the program. See Appendix K for the descriptive statistics for writing processes questionnaire items by group and time. For reliability analysis, the alpha for all scales was above .60 (see Appendix L for Cronbach"s Alpha results). However, there was one item whose deletion resulted in an increase in alpha, and its corrected correlation with the total was less than .30 for both pretest and posttests: Q23 (I changed my writing plan (e.g., structure and content)) from the conceptualization phase. Therefore, the item was dropped from further analysis increasing alpha for the conceptualization scale to .76 from .74 for the pretest and .75 from .73 for the posttest. Because the items had high unidimensionality, each of the phases was converted into a scale to reduce the variables to limit type 1 error. Table 4.2 summarizes the descriptive statistics for the writing process scale by group and time.

TABLE 4.2

Descriptive Statistics for the Writing Process Scales by Time and Group

	Experimental				**Comparison**			
	Pretest (Week 1)		**Posttest (Week 8)**		**Pretest (Week 1)**		**Posttest (Week 8)**	
	M	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**
Conceptualization Scale	2.88	0.48	3.30	0.39	3.10	0.23	3.42	0.26
Generating Ideas Scale	2.92	0.60	3.37	0.50	3.26	0.29	3.36	0.27
Organizing Ideas Scale	2.80	0.53	3.21	0.51	3.31	0.32	3.45	0.34
Generating Texts Scale	2.74	0.62	3.31	0.62	3.33	0.43	3.39	0.35
Monitoring and Revising at High-Level Scale	2.25	0.82	3.28	0.53	2.96	0.47	3.18	0.44
Monitoring and Revising at Low-Level Scale	2.04	0.99	3.31	0.46	2.80	0.55	2.96	0.42
Total for Time								
					Pretest		**Posttest**	
					M	**SD**	**M**	**SD**
Conceptualization Scale					2.98	0.39	3.36	0.33
Generating Ideas Scale					3.09	0.50	3.36	0.40
Organizing Ideas Scale					3.05	0.51	3.33	0.45
Generating Texts Scale					3.02	0.60	3.35	0.50
Monitoring and Revising at High-Level Scale					2.59	0.76	3.23	0.48
Monitoring and Revising at Low-Level Scale					2.41	0.88	3.14	0.47
Total for Group								
					Experimental		Comparison	
					M	SD	M	SD
Conceptualization Scale					3.09	0.48	3.26	0.29
Generating Ideas Scale					3.14	0.59	3.31	0.28
Organizing Ideas Scale					3.01	0.56	3.38	0.33
Generating Texts Scale					3.02	0.68	3.36	0.39
Monitoring and Revising at High-Level Scale					2.76	0.86	3.07	0.46
Monitoring and Revising at Low-Level Scale					2.67	1.00	2.88	0.49

N=16 for comparison and 17 for experimental group

Results of mixed MANOVA for the effects of time and group on the six writing processes scales found that there was a significant effect for time: V = 0.73, F(6, 26) = 11.63, p =

<.001, η2 = .73 and interaction effect for group with time: V = 0.57, F(6, 26) = 5.67, p = <.001, η2 = .57. The effect size shows that time had a greater effect than interaction effect for group with time. There was no significant effect for group.

Using a Bonferroni correction, follow-up ANOVA detected significant time effects for all scales as follows:

Conceptualization: F(1, 31), = 2.28, p = <.001, η2 = .57
Generating Ideas: F(1, 31), = 1.23, p = <.001, η2 = .36
Organizing Ideas: F(1, 31), = 1.24, p = <.001, η2 = .38
Generating Texts: F(1, 31), = 1.67, p = <.001, η2 = .44
Monitoring and Revising at High-Level: F(1, 31), = 6.49, p = <.001, η2 = .57
Monitoring and Revising at low-Level: F(1, 31), = 8.46, p = <.001, η2 = .51

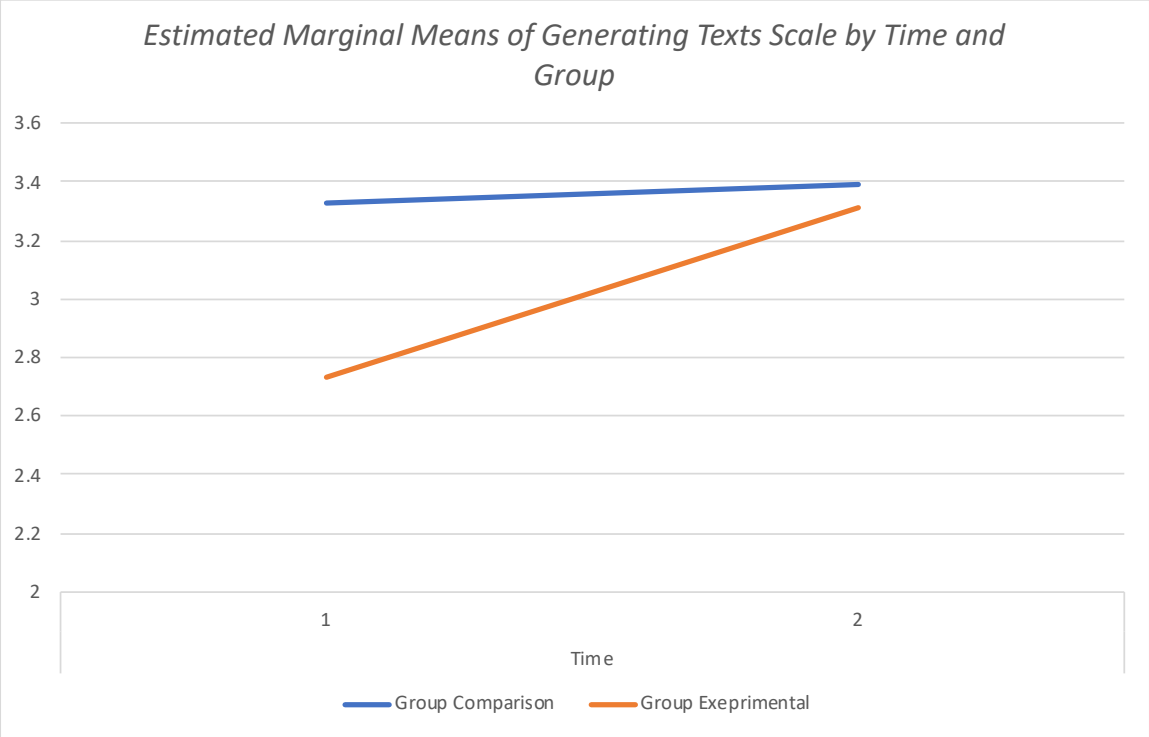
For all six scales, the mean was significantly higher for the posttest than the pretest for both groups, suggesting that the students reported engaging in each of the six processes more frequently at the end of the course than they did at the beginning of the course. On average, for each question, the mean for the posttest was higher than the pretest regardless of group.

ANOVA also detected significant interaction effects for group with time for generating texts: F(1, 31), = 1.67, p = <.001, η2 = .44, monitoring and revising at high-level: F(1, 31), = 6.49, p = <.001, η2 = .57, and monitoring and revising at low-level: F(1, 31), = 8.46, p = <.001, η2 = .51. Figures 4.1, 4.2, and 4.3 show the plots for the estimated marginal means for generating texts, monitoring and revising at high-level, and monitoring and revising at low-level scales by time and group.

Generating text: Figure 4.1 shows that the experimental group"s reported use of monitoring and revising at high-level was lower (M=2.74) than that for the comparison group (M=3.33) in the pretest. However, the increase in the reported use of generating

text for the experimental group was much higher, with both groups reporting similar levels of use by the posttest with an average increase of reported use of generating text from 2.74 to 3.31 for the experimental group and from 3.33 to 3.39 for the comparison group.

FIGURE 4.1
Estimated Marginal Means of Generating Texts by Time and Group



To examine the changes more in-depth, the means of scores for each question in the generating texts phase for pretest and posttest for each group were compared. Table 4.3 presents the descriptive statistics for generating texts phase items by group and time. While both groups reported thinking about correct sentence structure, correct grammar, and organizing sentences and paragraphs (Q16, Q17, and Q20), the experimental group exhibited a larger increase of about .65 or higher compared to .18 or lower for the comparison group for each question. For thinking about correct grammar (Q17), the students in the comparison group reported thinking less: the score for the comparison group decreased from 3.56 to 3.50, while the experimental group reported an increase from 2.76 to 3.41.

TABLE 4.3
Descriptive Statistics for Generating Texts Phase Items by Time and Group

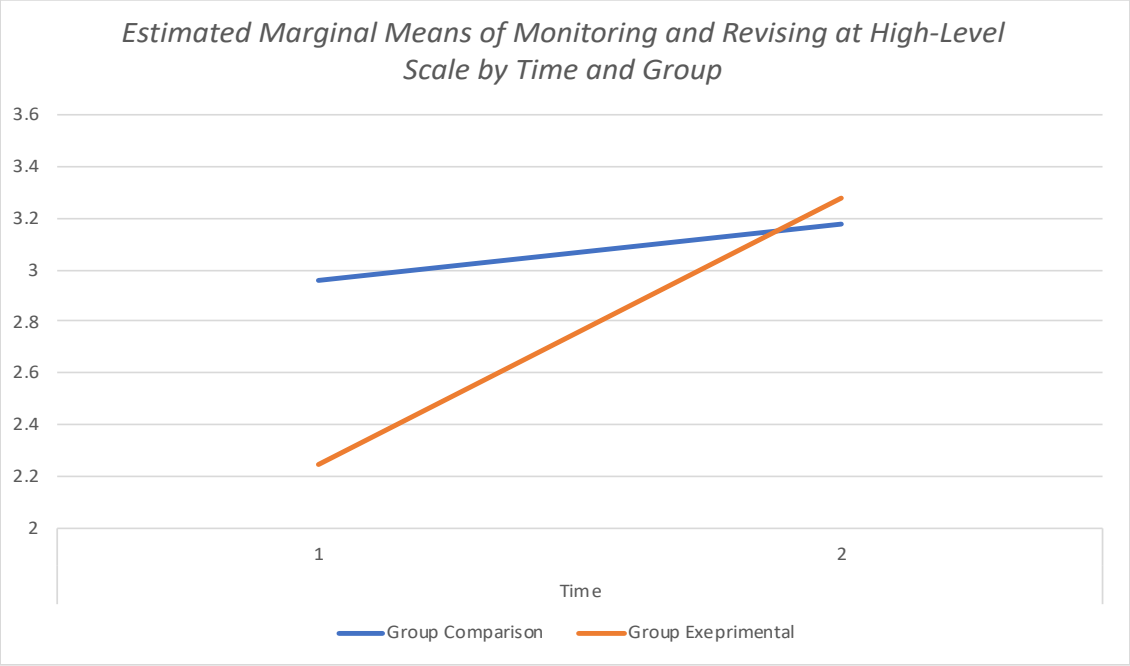
	Experimental				**Comparison**				**Change in Scores***	
	Pretest		**Posttest**		**Pretest**		**Posttest**		**Exp.**	**Comp.**
	M	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**		
Generating Texts Phase										
16. I thought about the correct sentence structures to express my ideas in the first draft	2.65	0.79	3.35	0.61	3.38	0.62	3.56	0.63	0.70	0.18
17. I thought about the correct grammar to express my ideas in the first draft	2.76	0.9	3.41	0.80	3.56	0.63	3.50	0.63	0.65	-0.06
18. I thought about how to connect my ideas smoothly in the whole essay in the first draft	2.94	0.83	3.24	0.75	3.19	0.40	3.25	0.45	0.30	0.06
20. I organized my sentences and paragraphs in a logical order in the first draft	2.59	0.62	3.24	0.56	3.19	0.54	3.25	0.45	0.65	0.06

N=16 for comparison and 17 for experimental group
* A negative value means a decrease between the preset and posttest.

Monitoring and revising at high level: Figure 4.2 shows that the experimental group"s reported use of monitoring and revising at high-level was lower (M=2.25) than that for the comparison group (M=2.96) in the pretest. However, by the posttest, the experimental group"s reported use increased to 3.28, while the comparison group only increased to 3.18.

FIGURE 4.2

Estimated Marginal Means of Monitoring and Revising at High-Level by Time and Group



To examine the changes more in-depth, the means of scores for each of the ten questions in the monitoring and revising at high-level phase for pretest and posttest for each group were compared. The ten questions in the Monitoring and Revising at High-level are divided into two parts: while writing the first draft (Q25 – Q29) and after receiving feedback (Q33- Q37). Table 4.4 presents descriptive statistics for these questions. While both groups reported checking organization, coherence, viewpoint, possible effect on the intended reader and accuracy and range of sentence structure in the first draft (Q25, Q26, Q27, Q28, and Q29), the experimental group exhibited an overall larger increase of at least .59 or higher compared to .25 or less for the comparison group for each question.

For items after receiving feedback (Q33-Q37), while both groups reported checking organization, coherence, viewpoint, possible effect on the intended reader and accuracy and range of sentence structure (Q33, Q34, Q35, Q36, and Q37), the experimental group exhibited a larger increase for each. For each question, the experimental group"s reported use of monitoring and revising at high-level increased, on average, by more than 1.18, while the comparison group increased by around .56 or less.

TABLE 4.4

Descriptive Statistics for the Monitoring and Revising at High-Level Items by Time and Group

	Experimental				**Comparison**				**Change in Scores**	
	Pretest		**Posttest**		**Pretest**		**Posttest**		**Exp.**	**Comp.**
	M	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**		
Monitoring and Revising at High-Level Phase										
25. I checked that my essay is well-organized and revised accordingly in the first draft	2.65	0.70	3.35	0.70	3.00	0.73	3.06	0.77	0.70	0.06
26. I checked that my essay is coherent and revised accordingly in the first draft	2.59	0.87	3.18	0.81	3.19	0.54	3.31	0.70	0.59	0.12
27. I checked that I include my own viewpoint on the topic and revised accordingly in the first draft	2.59	0.80	3.29	0.59	3.19	0.66	3.44	0.73	0.70	0.25
28. I checked the possible effect of my essay on the intended reader and revised accordingly in the first draft	2.41	0.71	3.12	0.60	3.31	0.60	3.38	0.62	0.71	0.07
29. I checked the accuracy and range of sentence structures and revised accordingly in the first draft	2.53	0.72	3.41	0.62	2.94	0.77	3.19	0.54	0.88	0.25
33. I checked that my essay was well organized and revised accordingly after receiving feedback	2.00	1.46	3.29	0.69	2.50	0.82	2.88	0.81	1.29	0.38
34. I checked that my essay was coherent and revised accordingly after receiving feedback	2.00	1.41	3.29	0.69	2.94	0.68	3.06	0.68	1.29	0.12
35. I checked that I include my own viewpoint on the topic and revised accordingly after receiving feedback	2.06	1.52	3.24	0.56	2.81	0.66	3.06	0.85	1.18	0.25
36. I checked the possible effect of my essay on the intended reader and revised accordingly after receiving feedback	1.76	1.35	3.18	0.64	2.94	0.68	3.13	0.81	1.42	0.19
37. I checked the accuracy and range of the sentence structures and revised accordingly after receiving feedback	1.88	1.36	3.41	0.62	2.75	0.68	3.31	0.70	1.53	0.56

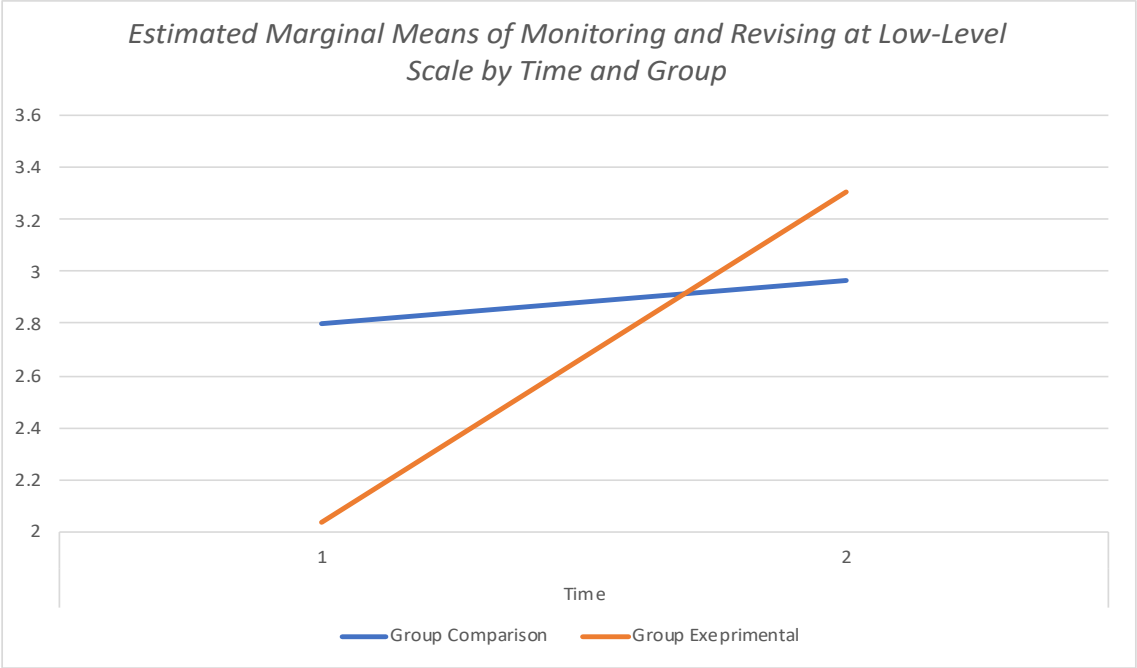
N=16 for comparison and 17 for experimental group

Monitoring and revising at low level: Figures 4.3 shows that the experimental group"s reported use of monitoring and revising at low-level was lower (M=2.04) than that for the

comparison group (M=2.80) in the pretest. However, by the posttest, the experimental group saw a significant increase in average reported use of monitoring and revising at low level from 2.04 to 3.31, while the comparison group only increased from 2.80 to 2.96.

FIGURE 4.3

Estimated Marginal Means of Monitoring and Revising at Low-Level by Time and Group



To examine the changes more in-depth, the means of scores for each question in the monitoring and revising at low-level phase for pretest and posttest for each group was compared. The questions in the Monitoring and Revising at Low-level are also divided into two parts: while writing the first draft (Q30-Q31) and after receiving feedback (Q32, Q38, and Q39), Table 4.5 presents them. While both groups reported checking grammar and the appropriateness and range of vocabulary in the first draft (Q30 and Q31), the experimental group exhibited a larger increase of about .5 or higher compared to .2 or less for the comparison group for each question.

For items after receiving feedback (Q32, 38, and 39), while both groups reported writing, checking grammar, and checking the appropriateness and range of vocabulary after receiving feedback (Q38, and 39), the experimental group exhibited a larger increase of

about 1.4 or higher compared to .4 or less for the comparison group for each question. For writing multiple drafts after feedback (Q32), the students in the comparison group reported writing fewer drafts after feedback: the score for the comparison group decreased from 2.81 to 2.69, while the experimental group reported an increase from 2.06 to 3.29.

TABLE 4.5
Descriptive Statistics for the Monitoring and Revising at Low-level Items by Time and Group

	Experimental				**Comparison**				**Change in Scores**	
	Pretest		**Posttest**		**Pretest**		**Posttest**		**Exp.**	**Comp.**
	M	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**		
Monitoring and Revising at Low-Level Phase										
30. I checked the grammar (e.g., part of speech and tenses) and revised accordingly in the first draft	2.82	0.73	3.29	0.85	3.06	0.77	3.25	0.68	0.47	0.19
31. I checked the appropriateness and range of vocabulary and revised accordingly in the first draft	1.65	0.79	3.35	0.49	2.75	1.00	2.88	0.81	1.70	0.13
32. I usually write multiple drafts after receiving feedback	2.06	1.48	3.29	0.69	2.81	0.75	2.69	0.87	1.23	-0.12
38. I checked the grammar (e.g., part of speech and tenses) and revised accordingly after receiving feedback	1.88	1.45	3.35	0.70	2.75	0.58	2.94	0.44	1.47	0.19
39. I checked the appropriateness and range of vocabulary and revised accordingly after receiving feedback	1.76	1.30	3.24	0.83	2.63	0.72	3.06	0.77	1.48	0.43

N=16 for comparison and 17 for experimental group

4.2 Changes in Writing Due to Hybrid Corrective Feedback

The following subsections report results concerning changes in fluency, syntactic complexity, linguistic accuracy, lexical complexity, organization, and task response to address research question 2.

4.2.1 Changes in Fluency

To examine changes in fluency, operationalized as the number of words per essay, across groups and time, a mixed ANOVA was conducted to examine if the changes in the number of words per essay for the experimental and comparison group across time were significant. Mauchly's sphericity test was nonsignificant (p = .26), so no corrections were made for the mixed ANOVA. Table 4.6 summarizes the descriptive statistics for fluency by group and time.

TABLE 4.6
Descriptive Statistics for Fluency by Group and Time

	Experimental		**Comparison**		**Total for Groups**	
	M	**SD**	**M**	**SD**	**M**	**SD**
Pretest (Week 1)	485.50	74.60	436.88	136.62	461.19	111.06
Writing Task 1 (Week 3)	560.94	106.38	565.19	106.29	563.06	104.63
Writing Task 2 (Week 5)	522.50	83.36	529.00	76.96	525.75	78.99
Writing Task 3 (Week 7)	581.75	72.00	650.00	110.06	615.88	97.84
Delayed Posttest (3 months after treatment)	517.13	90.59	477.25	81.66	497.19	87.22
Total Across Time	533.56	90.76	531.66	125.96		

N=16 for both comparison and experimental groups

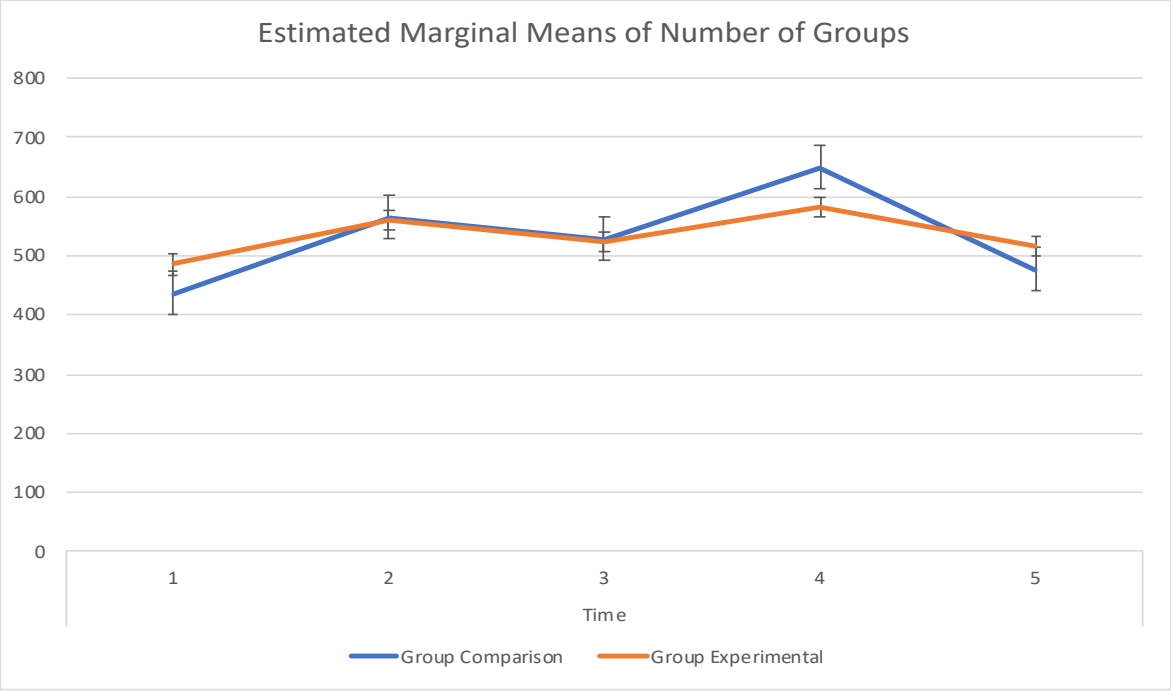
Fluency by groups and time: The analyses indicated that there was no significant effect for group but there was a significant effect for time: F(4, 120), = 22.65, p = .000, η2 = .43, and a significant effect was found for the interaction of group with time: F(4, 120), = 3.44, p = .011, η2 = .10. This significant interaction suggests that the length of the essay changed differently for the two groups across time. The effect size for interaction of group with time was lower than the effect size for time (.10 compared to .43). Figure 4.4 shows the plot for the estimated marginal means of number of words at five-time points by group. The plot indicates that the two groups followed a similar trend overall, but the comparison group showed larger differences across time points than did the experimental group. As Table 4.6 shows, essays written later tended to be longer than essays written earlier in the course. The mean essay length increased from 485.50 words in the pretest (week 1) to 581.75 words in writing task 3 (week 7) for the experimental group. However, the mean essay length for the comparison group increased more (from

M=436.88 in writing task 1 to M=650.00 in writing task 3). For the delayed posttest (3 months after treatment), while the difference between writing task 3 (week 7) and the delayed posttest (3 months after treatment) remained similar, there was a precipitous fall for the comparison group. For both groups, the number of words steadily increased from the pretest (week 1) to writing task 3 (week 7), seeing an increase of almost 150 words. However, there was a significant drop in the delayed posttest (3 months after treatment) to 497.19 words.

The analyses indicated that there was a significant effect for time: F(4, 120), = 22.65, p = .000, η2 = .43. As can be seen in Table 4.6, essays written later tended to be longer than essays written earlier in the course. Mean essay length increased from 485.50 words in week 1 to 581.75 words in week 7 for the experimental group, from 436.88 words to 650.00 words for the comparison group, and 461.19 to 497.19 for both. For both groups, the number of words steadily increased from pretest to writing task 3, seeing an increase of almost 150 words. However, there was a significant drop in the delayed posttest to 497.19 words.

FIGURE 4.4

Estimated Marginal Means for Fluency by Group and Time



Fluency by drafts and time: For changes in fluency across drafts and time for the experimental group, the analysis found no main or interaction effects for the covariate, number of drafts. Therefore, the analysis was done again without the covariate to simplify the interpretation. Mauchly"s test statistic was nonsignificant for time (p = .57) and for the interaction between time and drafts (p = .22). The analyses indicated that there was no significant effect for time or the interaction of draft with time. However, a significant effect was found for draft: F(1, 16), = 17.5., p = < .001, η2 = .52. As Table 4.7 shows, the length of essays did not vary much across time points for both the first and the last drafts. The first draft essays in writing task 1 (week 3) were on average 558.00 words, while essays written on writing task 3 (week 7) were 579.47 words long, on average. This was true for the last draft too. That is, the length of the last drafts did not vary much across time points: for both groups, the increase was marginal (M=579.79 for writing task 1 (week 3) and M=622.50 for writing task 3). On Average, the last draft had a significantly higher number of words (M=632.57) than did the initial draft (M=552.73). As Table 4.7 shows, there was a significant increase in the number of words between drafts at each time point. For example, in writing task 1 (week 3), the essays' length increased from 558.00 words for draft 1 to 601.59 words in the last draft, on average. Similarly, in writing task 3 (week 7), the essays' length increased from 579.47 words to 665.53 words, on average. In general, the students in the experimental group tended to write more when they rewrite their essays.

TABLE 4.7
Descriptive Statistics for Fluency by Draft and Time for the Experimental Group

	First Draft		Last Draft		Total Across Drafts	
	M	SD	M	SD	M	SD
Writing Task 1 (Week 3)	558	103.714	601.59	80.254	579.79	93.95
Writing Task 2 (Week 5)	520.71	81.054	630.59	132.221	575.65	121.54
Writing Task 3 (Week 7)	579.47	70.341	665.53	100.886	622.50	96.13
Total Across Time	552.73	87.92	632.57	107.75		

N=17 students.

4.2.2 Changes in Syntactic Complexity

The following indices were used for syntactic complexity: global complexity and dependent types of noun phrases. Global complexity: For global complexity, the following indices were used: Length of production unit - mean length of T-units (MLT), sentence complexity - sentence complexity ratio (C/S), amount of subordination - dependent clauses per T-unit (DC/T), amount of coordination - coordinate phrases per T-unit (CP/T), and degree of phrasal sophistication - complex nominals per T-unit (CN/T).

For changes in global complexity across groups and time, a mixed MANOVA indicated that there was a significant effect for group: V = .47, F(5, 26) = 4.52, p = .004, η2 = .465; time: V = .849, F(20, 11) = 3.08, p = .029, η2 = .849; and group with time: V = 0.83, F(20, 11) = 2.75, p = .044, η2 = .83. Table 4.8 summarizes the descriptive statistics for global complexity by group and time.

TABLE 4.8
Descriptive Statistics for Global Complexity by Group and Time

		Experimental		**Comparison**		**Total for Groups**	
		M	**SD**	**M**	**SD**	**M**	**SD**
Mean Length of T-Units (MLT)	Pretest (Week 1)	14.63	2.45	14.98	2.54	14.80	2.46
	Writing Task 1 (Week 3)	15.98	3.28	16.51	3.06	16.24	3.13
	Writing Task 2 (Week 5)	15.93	2.71	16.47	2.45	16.20	2.56
	Writing Task 3 (Week 7)	15.88	3.1	17.16	3.23	16.52	3.18
	Delayed Posttest (3 months after treatment)	16.92	3.52	15.04	1.72	15.98	2.89
	Total Across Time	15.87	3.05	16.03	2.73		
Sentence Complexity Ratio (C/S)	Pretest (Week 1)	2.04	0.52	2.12	0.34	2.08	0.44
	Writing Task 1 (Week 3)	2.01	0.57	2.36	0.67	2.19	0.63
	Writing Task 2 (Week 5)	1.92	0.31	2.18	0.47	2.05	0.41
	Writing Task 3 (Week 7)	2.3	0.45	2.38	0.55	2.34	0.50
	Delayed Posttest (3 months after treatment)	2.21	0.53	2.35	0.52	2.28	0.52
	Total Across Time	2.10	0.49	2.28	0.52		

		Experimental		Comparison		Total for Groups	
		M	**SD**	**M**	**SD**	**M**	**SD**
Dependent Clauses Per T-Unit (DC/T)	Pretest (Week 1)	0.61	0.22	0.72	0.23	0.66	0.23
	Writing Task 1 (Week 3)	0.59	0.25	0.68	0.2	0.63	0.23
	Writing Task 2 (Week 5)	0.57	0.19	0.73	0.15	0.65	0.19
	Writing Task 3 (Week 7)	0.71	0.24	0.82	0.3	0.76	0.27
	Delayed Posttest (3 months after treatment)	0.73	0.3	0.74	0.17	0.74	0.24
	Total Across Time	0.64	0.24	0.74	0.22		
Coordinate Phrases Per T-Unit (CP/T)	Pretest (Week 1)	0.36	0.12	0.33	0.15	0.34	0.13
	Writing Task 1 (Week 3)	0.39	0.15	0.46	0.19	0.43	0.17
	Writing Task 2 (Week 5)	0.35	0.14	0.33	0.13	0.34	0.13
	Writing Task 3 (Week 7)	0.36	0.17	0.36	0.2	0.36	0.18
	Delayed Posttest (3 months after treatment)	0.28	0.13	0.28	0.11	0.28	0.12
	Total Across Time	0.35	0.14	0.35	0.17		
Complex Nominals Per T-Unit (CN/T)	Pretest (Week 1)	1.71	0.32	1.65	0.39	1.68	0.35
	Writing Task 1 (Week 3)	1.92	0.43	1.8	0.39	1.86	0.40
	Writing Task 2 (Week 5)	2.27	0.5	2.09	0.46	2.18	0.48
	Writing Task 3 (Week 7)	2.16	0.52	2.06	0.42	2.11	0.47
	Delayed Posttest (3 months after treatment)	2.03	0.36	1.57	0.31	1.80	0.41
	Total Across Time	2.02	0.46	1.83	0.44		

N=16 for both comparison and experimental groups

Global complexity by groups and time: The follow-up univariate test for time showed significant effect for time at the Bonferroni corrected p-value (.01) for sentence complexity ratio, coordinate phrases per t-unit, and complex nominals per t-unit. Time did not significantly affect the mean length of t-units and dependent clauses per t-unit. There were no significant effects for group and the interaction effect for group with time on any of the indices at the Bonferroni corrected p-value (.01). The results for significant effect for time for sentence complexity ratio were $F(4, 120), = 3.73$, $p = .007$, $\eta 2 = .11$. As shown in Table 4.8, essays written later tended to include more clauses per sentence than essays written earlier in the course. The mean essay sentence complexity ratio increased from 2.04 in pretest (week 1) to 2.30 in writing task 3 (week 7) for the experimental group and from 2.12 to 2.38 for the comparison group. For both groups, by the posttest, essays had a higher mean sentence complexity ratio (M= 2.28) than they did

in the pretest (M= 2.08).

Coordinate phrases per t-unit was significantly different for time too: $F(4, 120), = 4.57$, $p = .002$, $\eta 2 = .13$. As can be seen in Table 4.8, coordinate phrases per T-unit fluctuated between time points. Mean essay coordinate phrases per t-unit increased from 0.36 in pretest (week 1) to 0.39 in writing task 1 (week 3) for the experimental group and from 0.33 to 0.46 for the comparison group. However, for both groups, coordinate phrases per t-unit decreased after writing task 1 (week 3). For both groups, essays in writing task 1 (week 3) had the highest mean coordinate phrases per t-unit (M=0.43), then decreased in writing task 2 (week 5) (M=0.34) and again in the posttest (M=0.28).

Complex nominals per t-unit also was significantly different for time: $F(4, 120), = 12.23$, $p = <.001$, $\eta 2 = .29$). As shown in Table 4.8, essays written later tended to include more complex nominals per T-unit than essays written earlier in the course. Mean essay complex nominals per t-unit increased from 1.71 in week 1 to 2.16 in writing task 3 (week 7) for the experimental group and from 1.65 to 2.06 for the comparison group. For both groups, later essays tended to have higher complex nominals per T-unit except for the posttest. Pretest had the lowest (M=1.68) with writing task 2 (week 5) and writing task 3 (week 7) having similar ratios (M=2.18 and M=2.11, respectively). While the sentence complexity ratio and the coordinate phrases per t-unit had similar effect sizes (.11 and .13, respectively), the larger effect size for complex nominals per t-unit (.29) shows that there was a larger magnitude of the difference across time for complex nominals per t-unit.

Global complexity by drafts and time: For changes in global complexity across drafts and time for the experimental group, the 2 X 3 MANOVA analyses indicated that there was a significant effect for time: $V = 0.88$, $F(10, 7) = 5.36$, $p = .018$, $\eta 2 = .88$. However, there were no significant effects for draft or for draft with time. Table 4.9 summarizes the descriptive statistics for global complexity by draft and time.

TABLE 4.9

Descriptive Statistics for Global Complexity by Draft and Time for the Experimental Group

		First Draft		**Last Draft**		**Total Across Drafts**	
		M	**SD**	**M**	**SD**	**M**	**SD**
Mean Length of T-Units (MLT)	Writing Task 1 (Week 3)	16.04	3.19	16.73	3.14	16.39	3.13
	Writing Task 2 (Week 5)	16.10	2.72	17.16	2.80	16.63	2.77
	Writing Task 3 (Week 7)	15.98	3.03	16.50	3.00	16.24	2.98
	Total Across Time	16.04	2.93	16.80	2.94		
Sentence Complexity Ratio (C/S)	Writing Task 1 (Week 3)	2.01	0.55	1.97	0.48	1.99	0.51
	Writing Task 2 (Week 5)	1.92	0.30	1.97	0.27	1.94	0.28
	Writing Task 3 (Week 7)	2.31	0.44	2.22	0.35	2.26	0.39
	Total Across Time	2.08	0.47	2.05	0.39		
Dependent Clauses Per T-Unit (DC/T)	Writing Task 1 (Week 3)	0.58	0.24	0.56	0.31	0.57	0.27
	Writing Task 2 (Week 5)	0.58	0.18	0.61	0.21	0.60	0.19
	Writing Task 3 (Week 7)	0.71	0.23	0.69	0.26	0.70	0.25
	Total Across Time	0.63	0.22	0.62	0.27		
Coordinate Phrases Per T-Unit (CP/T)	Writing Task 1 (Week 3)	0.39	0.15	0.46	0.12	0.43	0.14
	Writing Task 2 (Week 5)	0.35	0.13	0.45	0.19	0.40	0.17
	Writing Task 3 (Week 7)	0.37	0.16	0.46	0.18	0.42	0.18
	Total Across Time	0.37	0.15	0.46	0.16		
Complex Nominals Per T-Unit (CN/T)	Writing Task 1 (Week 3)	1.94	0.43	2.15	0.49	2.05	0.47
	Writing Task 2 (Week 5)	2.29	0.49	2.43	0.38	2.36	0.44
	Writing Task 3 (Week 7)	2.16	0.50	2.25	0.48	2.20	0.49
	Total Across Time	2.13	0.49	2.28	0.46		

N=17 students

The follow-up univariate test for time showed significant effect at the Bonferroni corrected p-value (.001) for sentence complexity ratio but not for the other indices. The results for sentence complexity ratio were $F(2, 32) = 6.95$, $p = .003$, $\eta 2 = .30$. As Tables 4.9 shows, for the experimental group, the sentence complexity ratio increased from 2.01 in week 2 to 2.31 in writing task 3 (week 7) for the first draft and from 1.97 to 2.22 for the last draft. For both drafts, sentence complexity ratio was highest in writing task 3 (week 7) (M=2.26).

Dependent types of noun phrases by groups and time: The following indices were used for dependent types of noun phrases: incidence of determiners, adjectival modifiers, prepositional phrases, and possessives. A mixed MANOVA analysis indicated that there was a significant effect for group: V = .44, F(4, 27) = 5.21, p = .003, η2 = .44 and for time: V = .93, F(16, 15) = 12.53, p = <.001, η2 = .93. However, there was no significant interaction effect for group with time. Table 4.10 summarizes the descriptive statistics for dependent types of noun phrases by group and time.

TABLE 4.10
Descriptive Statistics for Dependent Types of Noun Phrases by Group and Time

		Experimental		**Comparison**		**Total for Groups**	
		M	**SD**	**M**	**SD**	**M**	**SD**
Determiners (Det)	Pretest (Week 1)	0.24	0.072	0.22	0.069	0.23	0.070
	Writing Task 1 (Week 3)	0.23	0.070	0.21	0.052	0.22	0.062
	Writing Task 2 (Week 5)	0.22	0.098	0.19	0.054	0.21	0.080
	Writing Task 3 (Week 7)	0.19	0.066	0.21	0.061	0.20	0.063
	Delayed Posttest (3 months after treatment)	0.29	0.084	0.24	0.068	0.27	0.081
	Total Across Time	0.24	0.084	0.21	0.062		
Adjectival Modifiers (A.Mod)	Pretest (Week 1)	0.24	0.062	0.19	0.061	0.21	0.066
	Writing Task 1 (Week 3)	0.24	0.051	0.18	0.048	0.21	0.057
	Writing Task 2 (Week 5)	0.31	0.045	0.25	0.049	0.28	0.056
	Writing Task 3 (Week 7)	0.27	0.054	0.21	0.041	0.24	0.054
	Delayed Posttest (3 months after treatment)	0.18	0.070	0.14	0.054	0.16	0.064
	Total Across Time	0.25	0.071	0.20	0.061		
Prepositional Phrases (Pre.Phr)	Pretest (Week 1)	0.11	0.038	0.11	0.049	0.11	0.043
	Writing Task 1 (Week 3)	0.12	0.040	0.11	0.034	0.12	0.036
	Writing Task 2 (Week 5)	0.13	0.037	0.10	0.030	0.12	0.036
	Writing Task 3 (Week 7)	0.12	0.051	0.14	0.041	0.13	0.047
	Delayed Posttest (3 months after treatment)	0.14	0.044	0.10	0.039	0.12	0.045
	Total Across Time	0.12	0.043	0.11	0.040		

		Experimental		Comparison		Total for Groups	
		M	**SD**	**M**	**SD**	**M**	**SD**
Possessives (Poss)	Pretest (Week 1)	0.06	0.034	0.07	0.037	0.068	0.036
	Writing Task 1 (Week 3)	0.07	0.039	0.09	0.021	0.077	0.033
	Writing Task 2 (Week 5)	0.07	0.041	0.07	0.032	0.066	0.036
	Writing Task 3 (Week 7)	0.06	0.036	0.06	0.028	0.063	0.032
	Delayed Posttest (3 months after treatment)	0.06	0.035	0.08	0.030	0.069	0.034
	Total Across Time	0.062	0.036	0.074	0.030		

N=16 for both comparison and experimental groups

The follow-up univariate test for group showed significant effect at the Bonferroni corrected p-value (.013) for adjectival modifiers, but not for determiners, prepositional phrases, and possessives. The results for adjectival modifiers were $F(1, 30), = 14.87$, $p = <.001$, $\eta 2 = .33$). As can be seen in Table 4.10, on average, the experimental group had higher incidents of adjectival modifiers (M=.25) than did the comparison group (M=.20).

The univariate test for time indicated that the incidences of adjectival modifiers and determiners were significantly different across time but not for prepositional phrases and possessives. The results for adjectival modifiers were $F(4, 120), = 34.40$, $p = <.001$, $\eta 2 = .53$). The incidence of adjectival modifiers increased from the pretest (week 1) to writing task 2 (week 5). The mean number of adjectival modifiers increased from 0.24 in writing task 1 (week 3) to 0.31 in writing task 2 (week 5) for the experimental group and from 0.19 to 0.25 for the comparison group. However, for both groups, the number of adjectival modifiers decreased after writing task 2 (week 5). On average, for both groups, the incidents of adjectival modifiers increased from pretest (week 1) (M=0.21) to writing task 2 (week 5) (M=0.28). However, it decreased in writing task 3 (week 7) (M=.24) and again in the delayed posttest (3 months after treatment) (M=0.16).

The univariate test for time showed significant effect for the incidence of determiners also: $F(4, 120), = 8.14$, $p = <.001$, $\eta 2 = .21$. As can be seen in Table 4.10, essays in the delayed posttest (3 months after treatment) contained more determiners than did the

essays in the pretest (week 1). The mean incidence of determiners increased from .24 in week 1 to .29 in the delayed posttest (3 months after treatment) for the experimental group and from .22 to .24 for the comparison group. The effect sizes for group on adjectival modifiers, time on adjectival modifiers, and time on determiners were all large, with the incidences of adjectival modifiers across time having the largest magnitude of the difference compared to incidence of determiners across time and adjectival modifiers across groups.

Dependent types of noun phrases by drafts and time for the experimental group: The 2 X 3 MANOVA analyses indicated that there was a significant effect for time: $V = 0.82$, $F(8, 9) = 5.05$, $p = .013$, $\eta 2 = .82$ and draft: $V = 0.52$, $F(4, 13) = 3.57$, $p = .036$, $\eta 2 = .52$. The effect size for time was higher than draft indicating that the magnitude of the difference for time was greater than that for draft. However, there was no interaction effect for draft with time. Table 4.11 summarizes the descriptive statistics for dependent types of noun phrases draft and time.

TABLE 4.11

Descriptive Statistics for Indices of Dependent Types of Noun Phrases by Draft and Time for the Experimental Group

		First Draft		Last Draft		Total Across Drafts	
		M	**SD**	**M**	**SD**	**M**	**SD**
Determiners (Det)	Writing Task 1 (Week 3)	0.23	0.068	0.24	0.052	0.24	0.060
	Writing Task 2 (Week 5)	0.23	0.096	0.23	0.066	0.23	0.081
	Writing Task 3 (Week 7)	0.19	0.064	0.20	0.064	0.20	0.064
	Total Across Time	0.22	0.078	0.22	0.062		
Adjectival Modifiers (A.Mod)	Writing Task 1 (Week 3)	0.24	0.051	0.26	0.076	0.25	0.064
	Writing Task 2 (Week 5)	0.32	0.046	0.33	0.051	0.32	0.048
	Writing Task 3 (Week 7)	0.27	0.052	0.29	0.072	0.28	0.063
	Total Across Time	0.28	0.058	0.29	0.073		
Prepositional Phrases (Pre.Phr)	Writing Task 1 (Week 3)	0.12	0.039	0.15	0.044	0.14	0.045
	Writing Task 2 (Week 5)	0.13	0.039	0.15	0.039	0.14	0.039
	Writing Task 3 (Week 7)	0.11	0.050	0.14	0.058	0.13	0.055
	Total Across Time	0.12	0.043	0.15	0.047		
Possessives (Poss)	Writing Task 1 (Week 3)	0.066	0.038	0.075	0.031	0.070	0.035
	Writing Task 2 (Week 5)	0.068	0.040	0.068	0.033	0.068	0.036
	Writing Task 3 (Week 7)	0.063	0.036	0.062	0.036	0.062	0.036
	Total Across Time	0.066	0.037	0.068	0.033		

N=17 students

The follow-up univariate test for time showed a significant effect at the Bonferroni corrected p-value (.013) for adjectival modifiers. There were no significant effects for time on determiners, prepositional phrases, and possessives. The results for adjectival modifiers were $F(2, 32), = 14.61$, $p = <.001$, $\eta 2 = .48$. As can be seen in Table 4.11, essays written later tended to have more adjectival modifiers than essays written earlier in the course; the incidence of adjectival modifiers increased from .26 in writing task 1 (week 3) to .29 in writing task 3 (week 7) for the experimental group and from .24 to .26 for the comparison group. On average, for both groups, the incidents of adjectival modifiers increased from writing task 1 (week 3) (M=0.25) to writing task 2 (week 5) (M=0.32). However, it decreased in writing task 3 (week 7) (M=.28).

The univariate test for draft showed significant effect only for prepositional phrases: F(1, 16), = 10.79, p = <.005, η2 = .40. On average, irrespective of time, the incidence of prepositional phrases increased by twenty percent between the first and the last drafts (M=0.12 and M=0.15, respectively). While other indices increased between the first and the last drafts, they were more marginal at around 5 percent.

In summary, for changes in global complexity across groups and time, there was a significant effect for group, time, and group with time. Specifically, there were increases for both groups for sentence complexity ratio, coordinate phrases per t-unit, and complex nominals per t-unit. For changes in global complexity across drafts and time for the experimental group, the analyses indicated a significant positive effect for time, but there were no significant effects for draft or for draft with time. Specifically, there was a significant increase in sentence complexity ratio but not for the other indices after instruction. In addition, for changes in dependent types of noun phrases across groups and time, there was a significant positive effect for group and for time, but not for group with time: there was a significant increase for adjectival modifiers for the experimental group compared to the comparison group. Furthermore, incidences of adjectival modifiers and determiners were increased significantly across time but not for prepositional phrases and possessives for both groups. For changes in dependent types of noun phrases across drafts and time for the experimental group, the analyses indicated a significant effect for time and draft, but there was no significant interaction effect for draft with time. In other words, dependent types of noun phrases increased between drafts and across time. Specifically, after the treatment, there was a significant increase for adjectival modifiers but not for determiners, prepositional phrases, and possessives. However, for between drafts for the experimental group, the results only showed a significant effect for prepositional phrases; the prepositional phrases showed a marked increase between drafts.

4.2.3 Changes in Linguistic Accuracy

Linguistic accuracy was operationalized by the number of errors identified by Criterion per 100 words.

Linguistic accuracy by groups and time: A mixed univariate analysis was used. Mauchly's Test of Sphericity test statistic was significant for time (p = .004). Field (2018) suggests that when Mauchly's test is significant, the Greenhouse Geisser correction should be used if Greenhouse Geisser estimate of sphericity is less than .75 (.72 in the results). Table 4.12 summarizes the descriptive statistics for the number of errors per 100 words by group and time.

TABLE 4.12
Descriptive Statistics for Number of Errors Per 100 Words by Group and Time

	Experimental		**Comparison**		**Total for Groups**	
	M	**SD**	**M**	**SD**	**M**	**SD**
Pretest	3.04	1.78	2.47	1.33	2.76	1.57
Task 1	4.64	1.70	3.22	1.36	3.93	1.68
Task 2	3.71	1.74	2.99	1.34	3.35	1.57
Task 3	3.62	1.40	2.65	0.86	3.13	1.24
Delayed posttest	2.34	2.01	2.44	1.47	2.39	1.74
Total across time	3.47	1.86	2.75	1.29		

N=16 for both comparison and experimental groups

The Greenhouse Geisser corrected test statistic indicated that there was a significant effect for time: F(2.88, 120), = 5.90, p = .001, η2 = .16. A higher number of errors per 100 indicates that the input text contains more frequent errors and is, thus, less grammatically accurate. As shown in Table 4.12, essays written later tended to have fewer errors than essays written earlier in the course. Mean errors per 100 words decreased from 3.22 errors per 100 words in writing task 1 (week 3) to 2.44 in the delayed posttest (3 months after treatment) for the comparison group and from 4.06 errors to 2.34 for the experimental group.

There was also a significant effect for group: F(1, 30), = 4.80, p = .036, η2 = .14. In

general, the experimental group made more errors per 100 words than did the comparison group (M=3.47 and M=2.75, respectively). The effect sizes for time and group were similar and large (.16 and .14, respectively). There was no significant interaction effect for group with time.

Linguistic accuracy by drafts and time for the experimental group: The initial 2 X 3 ANCOVA analysis with the number of drafts as a covariate found effects for time but not for the interaction for time with the covariate. Therefore, the analysis was done again without the covariate to simplify the interpretation. Mauchly"s test statistic was nonsignificant for time (p = .26) and for time with draft (p = .18). Table 4.13 summarizes the descriptive statistics for the number of errors per 100 words by draft and time.

TABLE 4.13
Descriptive Statistics for Number of Errors Per 100 Words by Draft and Time

	First Draft		Last Draft		Total Across Drafts	
	M	SD	M	SD	M	SD
Writing Task 1 (Week 3)	4.53	1.70	1.68	1.79	3.10	2.25
Writing Task 2 (Week 5)	3.70	1.69	1.01	0.82	2.35	1.89
Writing Task 3 (Week 7)	3.56	1.37	1.17	1.18	2.37	1.75
Total Across Time	3.93	1.62	1.29	1.33		

N=17 students

The analysis found that there was a significant effect for draft: F(1, 16), = 74.90, p = < .001, η2 = .82. As seen in Table 4.13, there was a significant decrease in the number of errors per 100 words between drafts at each time point: on average, the first draft had 3.93 errors per 100 words, but there were 1.29 errors per 100 words for the last draft. Although descriptive statistics indicated a general decrease in errors in later drafts, the univariate analysis indicated that there was no significant effect for time or for the interaction of draft with time.

4.2.4 Changes in Lexical Complexity

For changes in lexical complexity, lexical frequency, lexical range, and lexical depth were examined.

Lexical frequency scores by groups and time: For lexical frequency, the following indices were used: frequency of all words, frequency of content words, and frequency of bigrams. A mixed MANOVA analysis was done and found a significant effect for time: V = .86, F(12, 19) = 9.59, p = < .001, η2 = .86. The large effect size shows that there was a large magnitude of difference across time. There was no significant effect for group or for group with time. Table 4.14 summarizes the descriptive statistics for lexical frequency by group and time.

TABLE 4.14
Descriptive Statistics for Lexical Frequency by Group and Time

		Experimental		**Comparison**		**Total for Groups**	
		M	**SD**	**M**	**SD**	**M**	**SD**
Frequency of All Words	Pretest (Week 1)	7125.44	1167.89	7075.09	1109.03	7076.54	1111.36
	Writing Task 1 (Week 3)	7207.58	1740.29	6615.04	1207.86	6912.64	1480.33
	Writing Task 2 (Week 5)	7311.66	1986.75	6225.48	912.79	6779.68	1593.68
	Writing Task 3 (Week 7)	6302.97	1455.10	6501.33	757.83	6416.46	1130.61
	Delayed Posttest (3 months after treatment)	8834.15	1886.01	7189.62	1286.36	8011.89	1794.37
	Total Across Time	7330.84	1788.90	6721.31	1106.18		
Frequency of Content Words	Pretest (Week 1)	1095.78	151.52	1159.04	126.41	1127.01	138.77
	Writing Task 1 (Week 3)	897.85	144.58	957.11	88.76	926.91	119.92
	Writing Task 2 (Week 5)	1000.83	171.70	1046.42	164.07	1027.56	165.74
	Writing Task 3 (Week 7)	999.78	114.94	1038.70	133.91	1018.53	122.45
	Delayed Posttest (3 months after treatment)	957.19	190.32	1060.88	132.20	1009.03	169.58
	Total Across Time	992.80	163.74	1052.43	143.41		
Frequency of Bigrams	Pretest (Week 1)	140.31	54.85	165.55	46.22	151.63	51.25
	Writing Task 1 (Week 3)	166.32	66.56	125.99	50.69	144.82	61.21
	Writing Task 2 (Week 5)	160.41	71.26	136.86	53.28	148.37	62.07
	Writing Task 3 (Week 7)	139.48	56.87	164.53	54.44	152.46	55.40
	Delayed Posttest (3 months after treatment)	221.65	54.49	165.18	71.00	193.41	68.55
	Total Across Time	163.92	65.88	151.62	56.92		

N=16 for both comparison and experimental groups

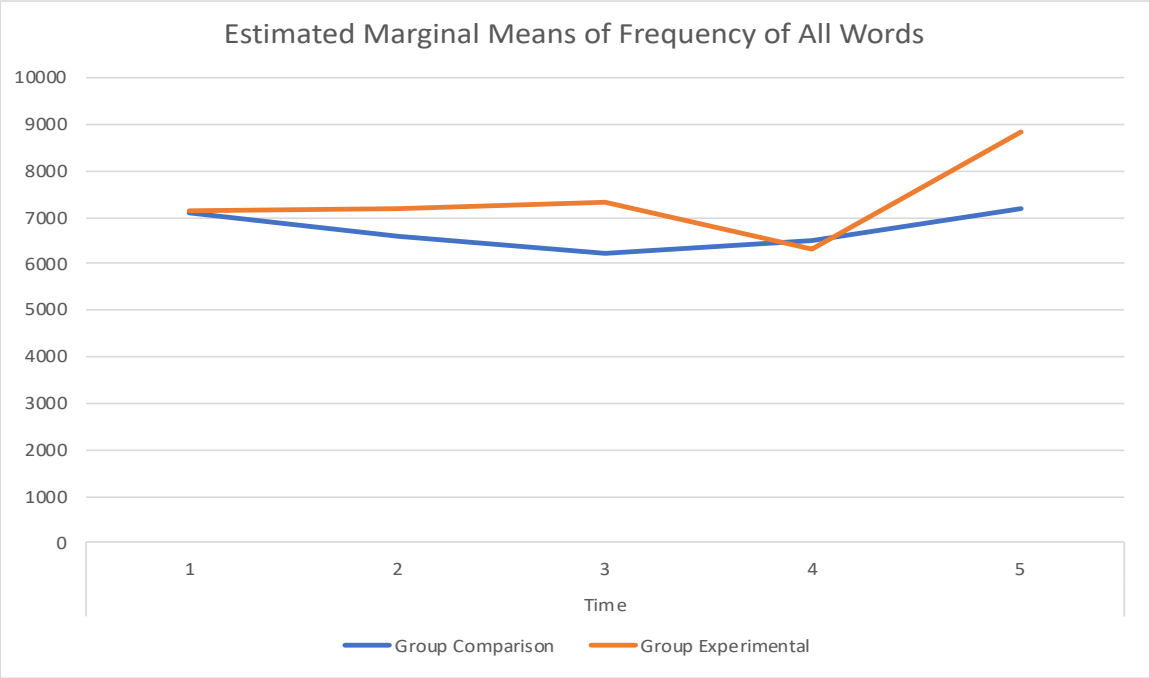
The follow-up univariate test for time showed significant effect at the Bonferroni corrected p-value (.017) for all three indices. The results for frequency of all words were F(4, 120), = 11.34, p = <.001, η2 = .28). Further examination revealed a significant

interaction effect for group with time for the univariate results (F(4, 120), = 4.45, p = <.002, η2 = .13) even though the multivariate results did not show a significant effect. A higher word frequency score indicates that the input text contains more frequent words and is, thus, less lexically sophisticated. On average, for both groups, frequency of all words decreased from the pretest (week 1) (M=7076.54) to writing task 3 (week 7) (M=6416.46) but saw a sharp increase in the delayed posttest (3 months after treatment) (M=8011.89). However, there was a wide margin of difference between the groups as seen in Table 4.14. Figure 4.5 shows the estimated marginal means of frequency by group and time.

The follow-up univariate test for time showed significant effect at the Bonferroni corrected p-value (.017) for all three indices. The results for frequency of all words scores were F(4, 120), = 11.34, p = <.001, η2 = .28). Further examination revealed a significant interaction effect for group with time for the univariate results (F(4, 120), = 4.45, p = <.002, η2 = .13) even though the multivariate results did not show a significant effect. Higher frequency scores indicate that the input text contains more frequent words and is, thus, less lexically sophisticated. On average, for both groups, frequency of all words scores decreased from the pretest (week 1) (M=7076.54) to writing task 3 (week 7) (M=6416.46) but saw a sharp increase in the delayed posttest (3 months after treatment) (M=8011.89). However, there was a wide margin of difference between the groups as seen in Table 4.14. Figure 4.5 shows the estimated marginal means of frequency of all words scores by group and time.

FIGURE 4.5

Estimated Marginal Means of Frequency of All Words by Group and Time



As can be seen in Figure 4.5, frequency of all words for both groups was similar in the pretest. while the experimental group steadily increased until writing task 2 (week 5) , the comparison group, in contrast, declined. While both groups saw an increase between writing task 3 (week 7) and the delayed posttest (3 months after treatment) , the increase for the experimental group was more marked. For example, the experimental group increased frequency of all words from 6302.97 to 8834.15 between writing task 3 (week 7) and the delayed posttest (3 months after treatment), the comparison group increased from 6501.33 to 7189.62.

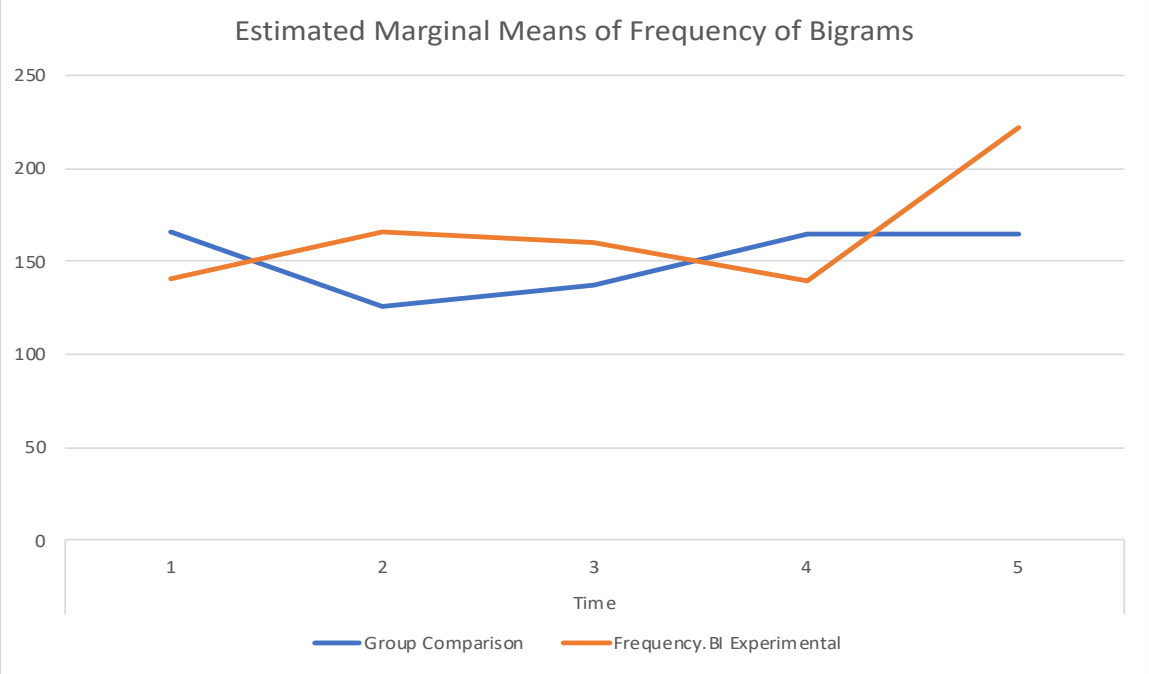
Frequency of content words was also significant for time: ($F(4, 120)$, = 10.62, $p = <.001$, $\eta 2 = .26$). The trends of change for both groups were similar with the comparison group consistently having a lower frequency, meaning the writing was more lexically sophisticated, with initial high frequency for both groups in the pretest (week 1) to a precipitous fall for writing task 1 (week 3): from 1095.78 on the pretest (week 1) to 897.85 on writing task 1 (week 3) for the experimental group and from 1159.04 to 957.11

for the comparison group. Then, there were gradual and similar increases for the experimental group until writing task 3 (week 7) reaching 999.78 and 1038.70 for the comparison group. The large effect size (η2 = .26) shows that there was a large magnitude of the difference across time for frequency of content words. It should be noted here that the differences in word frequencies across time might be due to differences in the topics of the tasks used in the study too.

Lastly, frequency of bigrams was also significant for time: (F(4, 120), = 4.45, p = .002, η2 = .13). However, there was a wide margin of difference between the groups as seen in Table 4.14. Further examination revealed a significant interaction effect for group with time for the univariate results (F(4, 120), = 4.18, p = <.003, η2 = .12) even though the multivariate results did not show a significant effect. The effect size for frequency of bigrams was low (η2 = .13). Figure 4.6 shows the estimated marginal means of frequency of bigrams by group and time.

FIGURE 4.6

Estimated Marginal Means of Bigram Frequency by Group and Time



As can be seen in Figure 4.6, the frequency of bigrams for the comparison group started higher but ended much lower than the experimental group. From the pretest (week 1) to writing task 3 (week 7), the two groups had an inverse trend: when the frequency for bigrams for the comparison group decreased, that of the experimental group increased. While both groups saw an increase between writing task 3 (week 7) and the delayed posttest (3 months after treatment), the increase for the experimental group was more apparent. For example, while the experimental group increased the frequency of bigrams from 139.48 to 221.65 between writing task 3 (week 7) and the delayed posttest (3 months after treatment), the comparison group only increased from 164.53 to 165.18. It should be noted here that the differences in bigram frequencies across time might be due to differences in the topics of the tasks used in the study too.

Lexical frequency by drafts and time for the experimental group: The 2 X 3 MANOVA analyses indicated that there was a significant effect for time: $V = 0.85$, $F(6, 11) = 10.32$, $p = .001$, $\eta 2 = .85$ but not for draft or draft with time. Table 4.15 summarizes the descriptive statistics for lexical frequency by draft and time.

TABLE 4.15
Descriptive Statistics for Lexical Frequency Scores by Draft and Time for the Experimental Group

		First Draft		**Last Draft**		**Total for Time**	
		M	**SD**	**M**	**SD**	**M**	**SD**
Frequency of All Words	Task 1	7192.73	1686.14	7618.03	1052.66	7405.38	1400.82
	Task 2	7301.27	1924.14	7713.73	1571.90	7507.50	1742.66
	Task 3	6336.58	1415.69	6832.60	1514.94	6584.59	1465.55
	Total for Draft	6943.53	1710.48	7388.12	1427.80		
Frequency of Content Words	Task 1	898.49	140.01	881.55	119.70	890.02	128.55
	Task 2	1009.82	170.32	983.00	147.77	996.41	157.60
	Task 3	999.55	111.30	1002.25	100.29	1000.90	104.33
	Total for Draft	969.28	148.64	955.60	132.86		
Frequency of Bigrams	Task 1	162.54	66.31	175.90	63.84	169.22	64.45
	Task 2	159.19	69.18	166.06	44.81	162.63	57.50
	Task 3	141.10	55.47	157.49	66.84	149.30	61.05
	Total for Draft	154.28	63.35	166.49	58.60		

N=17 students

The follow-up univariate test for time showed significant effect at the Bonferroni corrected p-value (.017) for frequency of all words and frequency of content words but not for bigram frequency. The results for frequency of all words were: $F(2, 32) = 6.69$, $p = .004$, $\eta 2 = .30$. Table 4.15 shows that the frequency of all words increased slightly from writing task 1 (week 3) (M=7405.38) to writing task 2 (week 5) (M=7507.50) but fell precipitously for writing task 3 (week 7) (M=6584.59). Again, this might be to topic effects as well.

In addition, frequency of content words also showed significant effect for time: $F(2, 32) = 7.61$, $p = .002$, $\eta 2 = .32$. On average, the frequency of content words increased from writing task 1 (week 3) to writing task 3 (week 7). For example, for the first drafts, the frequency of content words rose from 898.49 to 999.55. For the last drafts, it increased from 881.55 to 1002.25. However, there was not much difference between drafts (< 3% between drafts for each time point). On average, for both groups, the frequency of content words increased from writing task 1 (week 3) (M=890.02) to writing task 3 (week 7) (M=1000.90). Both the frequency of content words and the frequency of all words had similar large effect sizes ($\eta 2 = .32$ and 30, respectively).

Lexical range by groups and time: The following indices were used for lexical range: range of all words, range of function words, and range of bigrams. A mixed MANOVA analysis detected a significant effect for time: $V = .87$, $F(12, 19) = 11.02$, $p = < .001$, $\eta 2 = .87$. There was no significant effect for group or for the interaction of group with time. Table 4.16 summarizes the descriptive statistics for between-groups analysis of lexical range.

TABLE 4.16
Descriptive Statistics for Lexical Range by Group and Time

		Experimental		Comparison		Total for Groups	
		M	**SD**	**M**	**SD**	**M**	**SD**
Range of All Words	Pretest (Week 1)	0.62	0.031	0.63	0.021	0.63	0.027
	Writing Task 1 (Week 3)	0.61	0.031	0.62	0.019	0.62	0.026
	Writing Task 2 (Week 5)	0.61	0.027	0.61	0.025	0.61	0.026
	Writing Task 3 (Week 7)	0.60	0.027	0.62	0.021	0.61	0.025
	Delayed Posttest (3 months after treatment)	0.59	0.046	0.59	0.028	0.59	0.038
	Total Across Time	0.61	0.035	0.61	0.027		
Range of Function Words	Pretest (Week 1)	0.92	0.034	0.91	0.041	0.92	0.038
	Writing Task 1 (Week 3)	0.92	0.029	0.91	0.023	0.91	0.026
	Writing Task 2 (Week 5)	0.94	0.020	0.93	0.020	0.94	0.021
	Writing Task 3 (Week 7)	0.94	0.014	0.94	0.012	0.94	0.013
	Delayed Posttest (3 months after treatment)	0.94	0.014	0.93	0.024	0.94	0.021
	Total Across Time	0.93	0.026	0.92	0.028		
Range of Bigrams	Pretest (Week 1)	0.12	0.021	0.13	0.015	0.13	0.019
	Writing Task 1 (Week 3)	0.12	0.021	0.12	0.015	0.12	0.018
	Writing Task 2 (Week 5)	0.12	0.019	0.12	0.012	0.12	0.016
	Writing Task 3 (Week 7)	0.12	0.018	0.12	0.016	0.12	0.017
	Delayed Posttest (3 months after treatment)	0.14	0.019	0.13	0.022	0.14	0.020
	Total Across Time	0.12	0.021	0.13	0.018		

N=16 for both comparison and experimental groups

The follow-up univariate test for time showed significant effects at the Bonferroni corrected p-value (.017) for all three indices. The results for range of all words were F(4, 120), = 18.51, p = <.001, η2 = .38). For both groups, the range of all words fell steadily. A higher lexical range score means that the given text contains a wider range of words and thus demonstrates greater lexical sophistication, and a lower score means less lexical sophistication. For example, as seen in Table 4.16, essays written later tended to have lower range of all words than did those written earlier in the course. The range of all words decreased from .62 in the pretest (week 1) to .59 in the delayed posttest (3 months after treatment) for the experimental group and from .63 to .59 for the comparison group. On average, for both groups, range of all words decreased from the pretest (week 1)

(M=0.63) to the delayed posttest (3 months after treatment) (M=0.59).

The range of function words was also significant for time: F(4, 120), = 10.82, p = <.001, η2 = .27). However, in contrast to the range of all words, the range of function words increased. For the experimental group, it rose from .92 to .94 from the pretest (week 1) to the delayed posttest (3 months after treatment), and a similar increase was seen for the comparison group, from .91 to .93. On average, for both groups, the range of function words increased from the pretest (week 1) (M=0.92) to the delayed posttest (3 months after treatment) (M=0.94).

The range of bigrams was also significant for time: F(4, 120), = 12.02, p = <.001, η2 = .29. The range of Bigrams fell for the experimental group from the pretest (week 1) to writing task 3 (week 7) then rose for the delayed posttest (3 months after treatment). The comparison group fell from the pretest (week 1) to writing task 1 (week 3) then steadily rose afterwards. On average, for both groups, the range of bigrams increased from the pretest (week 1) (M=0.13) to the delayed posttest (3 months after treatment) (M=0.14). The analysis showed that all three indices had large effect sizes indicating that there was a large magnitude of the differences of indices for time.

The lexical range by drafts and time for the experimental group: The 2 X 3 MANOVA analyses indicated that there was a significant effect for time: V = 0.48, F(6, 62) = 3.29, p = .007, η2 = .24 and draft: V = 0.69, F(3, 14) = 10.25, p = <.001, η2 = .69. However, there was no significance for the interaction effect for draft with time. Table 4.17, summarizes the descriptive statistics for lexical range by draft and time.

TABLE 4.17
Descriptive Statistics for Lexical Range by Draft and Time for the Experimental Group

		First Draft		**Last Draft**		**Total for Time**	
		M	**SD**	**M**	**SD**	**M**	**SD**
Range AW	Task 1	0.61	0.031	0.61	0.022	0.61	0.026
	Task 2	0.61	0.027	0.60	0.025	0.60	0.026
	Task 3	0.60	0.026	0.60	0.025	0.60	0.025
	Total for Draft	0.61	0.028	0.60	0.024		
Range FW	Task 1	0.92	0.029	0.92	0.025	0.92	0.027
	Task 2	0.94	0.019	0.94	0.017	0.94	0.018
	Task 3	0.94	0.014	0.94	0.013	0.94	0.013
	Total for Draft	0.93	0.024	0.94	0.020		
Range BI	Task 1	0.12	0.021	0.12	0.017	0.12	0.019
	Task 2	0.12	0.019	0.12	0.013	0.12	0.016
	Task 3	0.12	0.017	0.12	0.017	0.12	0.017
	Total for Draft	0.12	0.019	0.12	0.016		

N=17 students

The follow-up univariate test for time showed significant effect at the Bonferroni corrected p-value (.017) for range of function words but not for the range of all words and bigram frequency. The results for range of function words were $F(2, 32) = 9.31$, $p = <.001$, $\eta 2 = .37$. As seen in Table 4.17, the range of function words increased from .92 to .94 for both groups, and there were no differences between the first and the last drafts.

The univariate test for draft showed significant effect only for range of all words: $F(1, 16), = 7.75$, $p = <.013$, $\eta 2 = .33$. On average, irrespective of time, the range of all words decreased slightly between the first and the last drafts (M=0.61 and M=0.60, respectively).

Lexical depth by groups and time: For lexical depth, the following indices were used: Academic Word List (AWL), Bigram Association Strength, Hypernymy Nouns, Hypernymy Verbs, Polysemy Adjectives, Polysemy Adverbs, Polysemy Content Words, and Contextual Distinctiveness. A mixed MANOVA analysis confirmed that there were no significant effects for time, group, or interaction effect for group with time. Table 4.18

summarizes the descriptive statistics for lexical depth by group and time.

TABLE 4.18
Descriptive Statistics for Lexical Depth by Group and Time

		Experimental		Comparison		Total for Groups	
		M	**SD**	**M**	**SD**	**M**	**SD**
Academic Word List (AWL)	Pretest (Week 1)	0.053	0.030	0.033	0.015	0.043	0.026
	Writing Task 1 (Week 3)	0.051	0.022	0.052	0.014	0.051	0.018
	Writing Task 2 (Week 5)	0.083	0.023	0.063	0.024	0.073	0.025
	Writing Task 3 (Week 7)	0.060	0.016	0.051	0.015	0.056	0.016
	Delayed Posttest (3 months after treatment)	0.049	0.021	0.047	0.023	0.048	0.022
	Total Across Time	0.059	0.026	0.049	0.021		
Bigram Association Strength (BAS)	Pretest (Week 1)	0.68	0.059	0.67	0.064	0.68	0.060
	Writing Task 1 (Week 3)	0.67	0.046	0.67	0.045	0.67	0.045
	Writing Task 2 (Week 5)	0.66	0.070	0.66	0.057	0.66	0.063
	Writing Task 3 (Week 7)	0.69	0.060	0.67	0.058	0.68	0.059
	Delayed Posttest (3 months after treatment)	0.67	0.063	0.66	0.049	0.67	0.056
	Total Across Time	0.67	0.059	0.67	0.054		
Hypernymy Nouns (Hyp.N)	Pretest (Week 1)	0.97	0.032	0.97	0.032	0.97	0.031
	Writing Task 1 (Week 3)	0.97	0.021	0.96	0.035	0.97	0.028
	Writing Task 2 (Week 5)	0.96	0.028	0.97	0.029	0.96	0.028
	Writing Task 3 (Week 7)	0.95	0.033	0.98	0.025	0.96	0.032
	Delayed Posttest (3 months after treatment)	0.96	0.032	0.96	0.024	0.96	0.028
	Total Across Time	0.96	0.030	0.97	0.029		
Hypernymy Verbs (Hyp.V)	Pretest (Week 1)	0.97	0.022	0.97	0.029	0.97	0.026
	Writing Task 1 (Week 3)	0.97	0.031	0.98	0.016	0.97	0.025
	Writing Task 2 (Week 5)	0.97	0.042	0.98	0.022	0.97	0.033
	Writing Task 3 (Week 7)	0.96	0.033	0.97	0.027	0.96	0.030
	Delayed Posttest (3 months after treatment)	0.98	0.027	0.98	0.027	0.98	0.027
	Total Across Time	0.97	0.032	0.97	0.025		
Polysemy Adjectives (Poly.Adj)	Pretest (Week 1)	0.94	0.089	0.95	0.091	0.95	0.089
	Writing Task 1 (Week 3)	0.96	0.077	0.94	0.082	0.95	0.079
	Writing Task 2 (Week 5)	0.98	0.051	0.96	0.087	0.97	0.071
	Writing Task 3 (Week 7)	0.95	0.079	0.95	0.074	0.95	0.075
	Delayed Posttest (3 months after treatment)	0.96	0.064	0.95	0.078	0.96	0.070
	Total Across Time	0.96	0.072	0.95	0.081		

		Experimental		Comparison		Total for Groups	
		M	**SD**	**M**	**SD**	**M**	**SD**
Polysemy Adverbs (Poly.Adv)	Pretest (Week 1)	0.97	0.055	0.96	0.054	0.96	0.054
	Writing Task 1 (Week 3)	0.97	0.063	0.97	0.050	0.97	0.056
	Writing Task 2 (Week 5)	0.93	0.109	0.94	0.107	0.93	0.106
	Writing Task 3 (Week 7)	0.98	0.036	0.97	0.033	0.98	0.034
	Delayed Posttest (3 months after treatment)	0.94	0.076	0.97	0.049	0.96	0.064
	Total Across Time	0.96	0.073	0.96	0.063		
Polysemy CW (Poly.CW)	Pretest (Week 1)	0.97	0.019	0.97	0.028	0.97	0.023
	Writing Task 1 (Week 3)	0.97	0.023	0.96	0.021	0.97	0.022
	Writing Task 2 (Week 5)	0.96	0.021	0.97	0.021	0.97	0.021
	Writing Task 3 (Week 7)	0.96	0.025	0.97	0.017	0.96	0.022
	Delayed Posttest (3 months after treatment)	0.97	0.025	0.97	0.018	0.97	0.021
	Total Across Time	0.96	0.023	0.97	0.021		
Contextual Distinctive-ness (CD)	Pretest (Week 1)	0.51	0.036	0.51	0.043	0.51	0.039
	Writing Task 1 (Week 3)	0.51	0.044	0.51	0.039	0.51	0.041
	Writing Task 2 (Week 5)	0.50	0.037	0.50	0.040	0.50	0.038
	Writing Task 3 (Week 7)	0.48	0.040	0.51	0.041	0.50	0.042
	Delayed Posttest (3 months after treatment)	0.50	0.050	0.50	0.034	0.50	0.042
	Total Across Time	0.50	0.042	0.50	0.039		

N=16 for both comparison and experimental groups

Lexical depth by drafts and time for the experimental group: The 2 X 3 MANOVA analyses indicated that there was a significant effect for draft only: $V = 0.81$, $F(8, 9) = 4.67$, $p = .017$, $\eta 2 = .81$. Table 4.19 summarizes the descriptive statistics for lexical depth by draft and time.

TABLE 4.19

Descriptive Statistics for Lexical Depth by Draft and Time for the Experimental Group

		First Draft		Last Draft		Total Across Drafts	
		M	**SD**	**M**	**SD**	**M**	**SD**
AWL value (AWL)	Writing Task 1 (Week 3)	0.052	0.022	0.066	0.018	0.059	0.021
	Writing Task 2 (Week 5)	0.083	0.022	0.093	0.026	0.088	0.024
	Writing Task 3 (Week 7)	0.059	0.016	0.067	0.013	0.063	0.015
	Total Across Time	0.065	0.024	0.075	0.023		
Bigram Association Strength (BAS)	Writing Task 1 (Week 3)	0.67	0.044	0.67	0.074	0.67	0.060
	Writing Task 2 (Week 5)	0.66	0.068	0.66	0.042	0.66	0.056
	Writing Task 3 (Week 7)	0.69	0.058	0.66	0.046	0.68	0.053
	Total Across Time	0.67	0.058	0.67	0.055		
Hypernymy Nouns (Hyp.N)	Writing Task 1 (Week 3)	0.97	0.022	0.96	0.035	0.97	0.029
	Writing Task 2 (Week 5)	0.96	0.029	0.96	0.029	0.96	0.028
	Writing Task 3 (Week 7)	0.95	0.035	0.97	0.024	0.96	0.031
	Total Across Time	0.96	0.029	0.96	0.029		
Hypernymy Verbs (Hyp.V)	Writing Task 1 (Week 3)	0.97	0.030	0.98	0.015	0.98	0.024
	Writing Task 2 (Week 5)	0.97	0.041	0.98	0.022	0.98	0.033
	Writing Task 3 (Week 7)	0.96	0.033	0.98	0.021	0.97	0.028
	Total Across Time	0.97	0.035	0.98	0.019		
Polysemy Adjectives (Poly.Adj)	Writing Task 1 (Week 3)	0.96	0.075	0.98	0.049	0.97	0.063
	Writing Task 2 (Week 5)	0.98	0.050	0.96	0.055	0.97	0.052
	Writing Task 3 (Week 7)	0.95	0.077	0.95	0.072	0.95	0.073
	Total Across Time	0.96	0.068	0.96	0.059		
Polysemy Adverbs (Poly.Adv)	Writing Task 1 (Week 3)	0.97	0.062	0.96	0.080	0.97	0.071
	Writing Task 2 (Week 5)	0.92	0.105	0.97	0.058	0.95	0.086
	Writing Task 3 (Week 7)	0.98	0.035	0.96	0.049	0.97	0.043
	Total Across Time	0.96	0.076	0.96	0.062		
Polysemy CW (Poly.CW)	Writing Task 1 (Week 3)	0.97	0.022	0.97	0.021	0.97	0.021
	Writing Task 2 (Week 5)	0.96	0.022	0.97	0.022	0.97	0.022
	Writing Task 3 (Week 7)	0.96	0.026	0.97	0.018	0.96	0.023
	Total Across Time	0.96	0.023	0.97	0.020		
Contextual Distinctiveness (CD)	Writing Task 1 (Week 3)	0.51	0.044	0.50	0.035	0.50	0.039
	Writing Task 2 (Week 5)	0.50	0.036	0.51	0.035	0.50	0.035
	Writing Task 3 (Week 7)	0.48	0.039	0.49	0.043	0.49	0.040
	Total Across Time	0.50	0.040	0.50	0.037		

N=17 students

The follow-up univariate test for draft showed significant effect at the Bonferroni corrected p-value (.006) for AWL but not for any of the other indices. The results for AWL were $F(1, 16) = 11.44$, $p = <.004$, $\eta 2 = .42$. As can be seen from Table 4.19, between the first and the last drafts for each writing task, there were significant increases. For writing task 1 (week 3), the increase was from 0.052 to 0.083; for writing task 2 (week 5), the increase was from 0.083 to 0.093; and for writing task 3 (week 7), the increase was from 0.059 to 0.067. On average, regardless of the task, the AWL value increased from the first draft (M=0.065) to the last draft (M=0.075). No other indices showed significant effect for draft.

In summary, for changes in lexical frequency scores across groups and time, there was a significant positive effect for time but not for group or for group with time. Specifically, there were increases for all three indices for both groups after the treatment, indicating the text contained more frequent words and is, thus, less lexically sophisticated: frequency of all words, content words, and bigrams. For changes in lexical frequency scores across drafts and time for the experimental group, the analyses indicated a significant effect for time but not for draft or draft with time. In detail, there were increases for frequency of all words and frequency of content words but not for bigram frequency after the treatment for both groups. In addition, for changes in lexical range across groups and time, the analysis found a significant effect for time but not for group or for the interaction of group with time; both groups saw increases after the treatment for all three indices: range of all words, range of function words, and range of bigrams. For changes in lexical range across drafts and time for the experimental group, the analyses indicated a significant effect for time and draft but not for the interaction effect for draft with time. For each new piece of writing, there was a significant increase in the range of function words but not for the range of all words and bigram frequency. However, for between drafts, the only significant increase was for the range of all words. Lastly, for changes in lexical depth across groups and time, the analysis confirmed that there were no significant effects for time, group, or interaction effect for group with time. For changes in lexical depth across drafts and time for the experimental group, the analyses

indicated a significant effect for draft but not for time or interaction effect for draft with time for the experimental group. Specifically, AWL value increased significantly between drafts for the experimental group. It should be noted here that some of the time effects on lexical features might be due to differences between the tasks the participants completed at different time points in terms of their topics and content. Unfortunately, given the design of the study, task and time effects are confounded.

4.2.5 Changes in Organization

For changes in organization, local cohesion, global cohesion, and text cohesion were examined.

Local cohesion by groups and time: For local cohesion, the following indices were used: adjacent sentence overlap for all words, adjacent sentence overlap for nouns, adjacent sentence overlap for verbs, and conceptual overlap of verbs between sentences. A mixed MANOVA analysis confirmed that there was a significant effect for group: $V = .29$, $F(4, 27) = 2.80$, $p = < .049$, $\eta 2 = .29$ and time: $V = .80$, $F(16, 15) = 3.77$, $p = .007$, $\eta 2 = .80$. The effect size for time was much larger (.80) than for group (.29) indicating that the magnitude of difference across time was larger than for group. However, there was no significant interaction effect for group with time. Table 4.20 summarizes the descriptive statistics for local cohesion by group and time.

TABLE 4.20

Descriptive Statistics for Local Cohesion by Group and Time

		Experimental		Comparison		Total for Groups	
		M	**SD**	**M**	**SD**	**M**	**SD**
Adjacent Sentence Overlap for All Words (ASO.AW)	Pretest (Week 1)	0.21	0.034	0.19	0.022	0.20	0.029
	Writing Task 1 (Week 3)	0.21	0.042	0.20	0.027	0.21	0.036
	Writing Task 2 (Week 5)	0.23	0.032	0.21	0.039	0.22	0.037
	Writing Task 3 (Week 7)	0.22	0.033	0.22	0.034	0.22	0.033
	Delayed Posttest (3 months after treatment)	0.22	0.043	0.21	0.031	0.22	0.038
	Total Across Time	0.22	0.037	0.21	0.032		
Adjacent Sentence Overlap for Nouns (ASO.N)	Pretest (Week 1)	0.19	0.059	0.15	0.062	0.17	0.063
	Writing Task 1 (Week 3)	0.19	0.046	0.14	0.053	0.17	0.054
	Writing Task 2 (Week 5)	0.22	0.061	0.20	0.047	0.21	0.056
	Writing Task 3 (Week 7)	0.22	0.056	0.21	0.057	0.22	0.056
	Delayed Posttest (3 months after treatment)	0.20	0.062	0.18	0.043	0.19	0.053
	Total Across Time	0.20	0.058	0.18	0.058		
Adjacent Sentence Overlap for Verbs (ASO.V)	Pretest (Week 1)	0.11	0.048	0.08	0.039	0.10	0.045
	Writing Task 1 (Week 3)	0.13	0.050	0.11	0.035	0.12	0.043
	Writing Task 2 (Week 5)	0.12	0.057	0.12	0.071	0.12	0.064
	Writing Task 3 (Week 7)	0.12	0.057	0.12	0.046	0.12	0.051
	Delayed Posttest (3 months after treatment)	0.13	0.058	0.10	0.048	0.11	0.054
	Total Across Time	0.12	0.053	0.11	0.050		
Conceptual Overlap of Verbs Between Sentences (CO.V)	Pretest (Week 1)	0.67	0.258	0.51	0.352	0.59	0.31
	Writing Task 1 (Week 3)	0.76	0.458	0.77	0.324	0.76	0.39
	Writing Task 2 (Week 5)	0.58	0.257	0.85	0.624	0.71	0.49
	Writing Task 3 (Week 7)	0.77	0.455	0.71	0.478	0.74	0.46
	Delayed Posttest (3 months after treatment)	0.73	0.379	0.78	0.478	0.76	0.43
	Total Across Time	0.70	0.370	0.72	0.467		

N=16 for both comparison and experimental groups

The follow-up univariate analysis found no significant effects for group for the Bonferroni corrected p-value (.013). The univariate analysis for time indicated there was significant effect for adjacent sentence overlap for all words and adjacent sentence overlap for all nouns. The results for adjacent sentence overlap for all words were F(4, 120), = 4.01, p = .001, $\eta 2$ = .14. As seen in Table 4.20, for both groups, adjacent sentence

overlap for all worlds increased. For example, incidents of adjacent sentence overlap for all words increased from .21 to .22 from the pretest (week 1) to the delayed posttest (3 months after treatment) for the experimental group and from .19 to .21 for the comparison group. On average, for both groups, the incidents of adjacent sentence overlap for all words increased from the pretest (week 1) (M=0.20) to the delayed posttest (3 months after treatment) (M=0.22).

In addition, the univariate analysis for time also found the change in adjacent sentence overlap for nouns to be significant. F(4, 120), = 8.57, p = <.001, η2 = .22). For both groups, adjacent sentence overlap for nouns increased. For example, adjacent sentence overlap for nouns increased from .19 to .20 from the pretest (week 1) to the delayed posttest (3 months after treatment) for the experimental group, and for the comparison group increased from .15 to .18. On average, for both groups, the incidents of adjacent sentence overlap for nouns increased from pretest (M=0.17) to delayed posttest (M=0.19). The greater effect size for adjacent sentence overlap for nouns compared to adjacent sentence overlap for all words indicates that the magnitude of the differences was larger for adjacent sentence overlap for nouns for time (η2 = .22 vs η2 = .14, respectively).

Local cohesion by drafts and time for the experimental group: The 2 X 3 MANOVA analyses indicated that there was a significant effect for draft (V = 0.50, F(4, 13) = 3.25, p = .047, η2 = .50) but not for time or draft with time. Table 4.21 summarizes the descriptive statistics for local cohesion by draft and time.

TABLE 4.21
Descriptive Statistics for Local Cohesion by Draft and Time for the Experimental Group

		First Draft		**Last Draft**		**Total Across Drafts**	
		M	**SD**	**M**	**SD**	**M**	**SD**
Adjacent Sentence Overlap for All Words (ASO.AW)	Writing Task 1 (Week 3)	0.21	0.041	0.22	0.048	0.22	0.044
	Writing Task 2 (Week 5)	0.23	0.032	0.22	0.030	0.22	0.031
	Writing Task 3 (Week 7)	0.22	0.033	0.22	0.028	0.22	0.030
	Total Across Time	0.22	0.035	0.22	0.036		
Adjacent Sentence Overlap for Nouns (ASO.N)	Writing Task 1 (Week 3)	0.19	0.045	0.19	0.053	0.19	0.048
	Writing Task 2 (Week 5)	0.22	0.060	0.18	0.050	0.20	0.058
	Writing Task 3 (Week 7)	0.23	0.059	0.21	0.051	0.22	0.055
	Total Across Time	0.21	0.056	0.19	0.051		
Adjacent Sentence Overlap for Verbs (ASO.V)	Writing Task 1 (Week 3)	0.13	0.051	0.11	0.047	0.12	0.049
	Writing Task 2 (Week 5)	0.11	0.060	0.09	0.049	0.10	0.055
	Writing Task 3 (Week 7)	0.12	0.055	0.12	0.050	0.12	0.052
	Total Across Time	0.12	0.055	0.11	0.049		
Conceptual Overlap of Verbs Between Sentences (CO.V)	Writing Task 1 (Week 3)	0.77	0.446	0.70	0.478	0.74	0.457
	Writing Task 2 (Week 5)	0.56	0.264	0.49	0.344	0.53	0.304
	Writing Task 3 (Week 7)	0.78	0.441	0.65	0.375	0.72	0.408
	Total Across Time	0.70	0.399	0.61	0.405		

N=17 students

The follow-up univariate test for draft showed no significant effect at the Bonferroni corrected p-value (.013).

Global cohesion by groups and time: A mixed MANOVA analysis confirmed that there

was no significant effect for group or for the interaction of group with time. However, there was a significant effect for time: V = .90, F(12, 19) = 14.38, p = < .001, η2 = .90. Table 4.22 summarizes the descriptive statistics for global cohesion by group and time.

TABLE 4.22
Descriptive Statistics for Global Cohesion by Group and Time

		Experimental		**Comparison**		**Total for Groups**	
		M	**SD**	**M**	**SD**	**M**	**SD**
Adjacent Paragraph Overlap for All Words (APO.AW)	Pretest (Week 1)	0.30	0.063	0.29	0.076	0.30	0.069
	Writing Task 1 (Week 3)	0.36	0.055	0.34	0.038	0.35	0.047
	Writing Task 2 (Week 5)	0.40	0.070	0.37	0.042	0.39	0.060
	Writing Task 3 (Week 7)	0.38	0.061	0.40	0.040	0.39	0.052
	Delayed Posttest (3 months after treatment)	0.36	0.075	0.36	0.046	0.36	0.062
	Total Across Time	0.36	0.072	0.35	0.060		
Conceptual Overlap of Verbs Between Paragraphs (CO.V)	Pretest (Week 1)	13.26	7.349	9.83	7.477	11.54	7.50
	Writing Task 1 (Week 3)	14.82	7.594	13.08	3.697	13.95	5.94
	Writing Task 2 (Week 5)	11.51	6.255	15.71	4.725	13.61	5.86
	Writing Task 3 (Week 7)	15.30	3.975	15.26	7.416	15.28	5.85
	Delayed Posttest (3 months after treatment)	13.81	7.114	14.04	6.176	13.92	6.55
	Total Across Time	13.74	6.56	13.59	6.29		
Conceptual Overlap of Nouns Between Paragraphs (CO.N)	Pretest (Week 1)	9.31	4.316	7.06	3.627	8.18	4.09
	Writing Task 1 (Week 3)	11.92	4.209	9.28	3.531	10.60	4.05
	Writing Task 2 (Week 5)	14.37	5.450	13.95	3.994	14.16	4.71
	Writing Task 3 (Week 7)	17.23	6.007	18.69	4.250	17.96	5.17
	Delayed Posttest (3 months after treatment)	15.49	6.655	12.54	3.727	14.01	5.51
	Total Across Time	13.66	5.97	12.30	5.50		

N=16 for both comparison and experimental groups

The follow-up univariate test for time showed significant effect at the Bonferroni corrected p-value (.017) for adjacent paragraph overlap for all words and conceptual overlap of nouns between paragraphs but not for other indices. The results for adjacent paragraph overlap for all words were F(4, 120), = 18.96, p = <.001, η2 = .39. Both groups increased adjacent paragraph overlap for all words until about the halfway point; then, there was a decline. For example, as seen in Table 4.22, incidents of adjacent paragraph overlap for all words increased from .30 to .40 from the pretest (week 1) to writing task 2

(week 5) for the experimental group, then decreased to .36 at the delayed posttest (3 months after treatment). Similarly, the comparison group increased from .29 to .40 from the pretest (week 1) to writing task 3 (week 7), then decreased to .36 in the delayed posttest (3 months after treatment). On average, for both groups, incidents of adjacent paragraph overlap for all words increased from the pretest (week 1) (M=0.30) to the delayed posttest (3 months after treatment) (M=0.36).

In addition, the univariate analysis for time also found conceptual overlap of nouns between paragraphs to be significant: F(4, 120), = 26.91, p = <.001, η2 = .47). Both groups' incidents of conceptual overlap of nouns between paragraphs increased from the pretest (week 1) to writing task 3 (week 7) (9.31 to 17.23 for the experimental group and 7.06 to 18.69 for the comparison group). Then, in the delayed posttest (3 months after treatment), the experimental group fell slightly to =15.49 while the comparison group fell to 12.54. On average, for both groups, the incidents of conceptual overlap of nouns increased from the pretest (week 1) (M=8.18) to writing task 2 (week 5) (M=14.01). The greater effect size for conceptual overlap of nouns between paragraphs compared to adjacent paragraph overlap for all words indicates that the magnitude of the differences was larger for conceptual overlap of nouns between paragraphs for time (η2 = .22 vs η2 = .14, respectively).

Global cohesion by drafts and time for the experimental group: The 2 X 3 MANOVA analyses indicated that there was a significant effect for time: V = 0.67, F(6, 11) = 3.70, p = .029, η2 = .67, draft: V = 0.59, F(3, 14) = 6.83, p = .005, η2 = .59, and draft with time V = 0.75, F(6, 11) = 5.52, p = .007, η2 = .75. The effect size for the interaction effect for draft with time was the largest indicating that it had the greatest magnitude of change compared to across time or for draft. Table 4.23 summarizes the descriptive statistics for global cohesion by draft and time.

TABLE 4.23
Descriptive Statistics for Global Cohesion by Draft and Time for the Experimental Group

		First Draft		Last Draft		Total Across Drafts	
		M	**SD**	**M**	**SD**	**M**	**SD**
Adjacent Paragraph Overlap for All Words (APO.AW)	Writing Task 1 (Week 3)	0.36	0.053	0.35	0.047	0.35	0.050
	Writing Task 2 (Week 5)	0.40	0.069	0.35	0.062	0.38	0.069
	Writing Task 3 (Week 7)	0.38	0.064	0.36	0.061	0.37	0.062
	Total Across Time	0.38	0.064	0.36	0.056		
Conceptual Overlap of Verbs Between Paragraphs (CO.V)	Writing Task 1 (Week 3)	14.50	7.47	12.82	5.81	13.66	6.64
	Writing Task 2 (Week 5)	11.21	6.18	12.89	10.16	12.05	8.33
	Writing Task 3 (Week 7)	15.81	4.39	13.62	4.51	14.72	4.52
	Total Across Time	13.84	6.33	13.11	7.11		
Conceptual Overlap of Nouns Between Paragraphs (CO.N)	Writing Task 1 (Week 3)	11.94	4.08	13.50	4.47	12.72	4.29
	Writing Task 2 (Week 5)	14.24	5.30	15.19	3.18	14.72	4.33
	Writing Task 3 (Week 7)	17.29	5.82	18.09	4.93	17.69	5.33
	Total Across Time	14.49	5.48	15.59	4.59		

N=17 students

The follow-up univariate test for time showed significant effect at the Bonferroni corrected p-value (.017) for only conceptual overlap of nouns between paragraphs: $F(2, 32) = 10.65$, $p = <.001$, $\eta 2 = .40$. As seen in Table 4.23, for both groups, there were large increases. For example, the first drafts indicated an increase in conceptual overlap of nouns between paragraphs from 11.94 to 17.29 between writing task 1 (week 3) and writing task 3 (week 7), while the last drafts increased from 13.50 to 18.09. On average, for both groups, the incidents of conceptual overlap of nouns between paragraphs increased from pretest (M=12.72) to task 2 (M=17.69). Although the multivariate test showed significant effect for draft and draft with time, due to the Bonferroni corrected p-value (.017), the univariate tests did not indicate significant effect for draft and for draft with time.

Text cohesion by groups and time: For text cohesion, the following indices were used: repeated content words (R.CW), pronoun density (Pron.Den), and all connectives (All.Con). A mixed MANOVA analysis confirmed that there was no significant

interaction effect for group with time, but there was a significant effect for time: V = .244, F(3, 28) = 3.00, p = < .047, η2 = .24 and group: V = .88, F(12, 19) = 11.15, p = < .001, η2 = .88. The larger effect size for group (.88) shows that there was a large magnitude of difference between groups for text cohesion indices compared to time. Table 4.24 summarizes the descriptive statistics for global cohesion by group and time.

TABLE 4.24
Descriptive Statistics for Text Cohesion by Group and Time

		Experimental		**Comparison**		**Total for Groups**	
		M	**SD**	**M**	**SD**	**M**	**SD**
Repeated Content Words (R.CW)	Pretest (Week 1)	0.37	0.061	0.33	0.047	0.35	0.058
	Writing Task 1 (Week 3)	0.41	0.041	0.36	0.045	0.38	0.048
	Writing Task 2 (Week 5)	0.45	0.068	0.42	0.045	0.43	0.059
	Writing Task 3 (Week 7)	0.45	0.063	0.43	0.036	0.44	0.053
	Delayed Posttest (3 months after treatment)	0.38	0.042	0.39	0.042	0.39	0.041
	Total Across Time	0.41	0.064	0.38	0.056		
Pronoun Density (Pron.Den)	Pretest (Week 1)	0.066	0.026	0.079	0.024	0.073	0.026
	Writing Task 1 (Week 3)	0.049	0.022	0.061	0.018	0.055	0.020
	Writing Task 2 (Week 5)	0.053	0.014	0.057	0.014	0.055	0.014
	Writing Task 3 (Week 7)	0.054	0.019	0.062	0.014	0.058	0.017
	Delayed Posttest (3 months after treatment)	0.049	0.021	0.069	0.017	0.059	0.021
	Total Across Time	0.054	0.021	0.066	0.019		
All Connectives (All.Con)	Pretest (Week 1)	0.077	0.014	0.077	0.015	0.077	0.014
	Writing Task 1 (Week 3)	0.081	0.014	0.083	0.011	0.082	0.012
	Writing Task 2 (Week 5)	0.069	0.014	0.077	0.016	0.073	0.015
	Writing Task 3 (Week 7)	0.078	0.016	0.078	0.012	0.078	0.014
	Delayed Posttest (3 months after treatment)	0.083	0.014	0.089	0.016	0.086	0.015
	Total Across Time	0.078	0.015	0.081	0.014		

N=16 for both comparison and experimental groups

The follow-up univariate analysis for group did not show any significant effect at the Bonferroni corrected p-value (.17). The univariate analysis for time indicated that all indices were significant. On average, the number of repeated content words increased (F(4, 120), = 36.41, p = <.001, η2 = .55). For example, as seen in Table 4.24, repeated content words increased from .37 to .45 between the pretest (week 1) and writing task 3

(week 7) for the experimental group, and the comparison group increased from .33 to .43 during the same period. However, both groups decreased in the delayed posttest (3 months after treatment): to .38 for the experimental and .39 for the comparison. On average, for both groups, the number of repeated content words increased steadily from the pretest (week 1) (M=0.35 to task 3 (M=0.44). However, it decreased in the delayed posttest (3 months after treatment) (M=.39).

Pronoun density was found to be significant for time: F(4, 120), = 7.02, p = <.001, η2 = .19. On average, both groups waned. The experimental group fell from .07 to .05 from the pretest (week 1) to the delayed posttest (3 months after treatment), while the comparison group fell from .08 to .07. On average, for both groups, pronoun density decreased from the pretest (week 1) (M=0.07) to the delayed posttest (3 months after treatment) (M=0.06).

All connectives were also found to be significant for time: F(4, 120), = 5.20, p = <.001, η2 = .15. Both groups increased the number of all connectives between the pretest (week 1) and posttests. The experimental group rose from .077 to .083 between the pretest to the delayed posttest (3 months after treatment), while the comparison group rose from .077 to .089. On average, for both groups, the incidents of all connectives decreased from the pretest (week 1) (M=0.077, SD=0.014) to the delayed posttest (3 months after treatment) (M=0.086). The greater effect size for repeated content words (η2 = .55) compared to pronoun density (η2 = .19) and all connectives (η2 = .15) indicate that the magnitude of the differences was larger for repeated content words for time.

Text cohesion by drafts and time for the experimental group: The 2 X 3 MANOVA analyses indicated that there was a significant effect for time (V = 0.82, F(6, 11) = 8.34, p = .001, η2 = .82 and draft: V = 0.59, F(3, 14) = 6.71, p = .005, η2 = .59) but not for draft with time. Table 4.25 summarizes the descriptive statistics for global cohesion by draft and time.

TABLE 4.25
Descriptive Statistics for Text Cohesion by Draft and Time for the Experimental Group

		First Draft		Last Draft		Total Across Drafts	
		M	**SD**	**M**	**SD**	**M**	**SD**
Repeated Content Words (R.CW)	Writing Task 1 (Week 3)	0.41	0.040	0.39	0.031	0.40	0.037
	Writing Task 2 (Week 5)	0.44	0.066	0.41	0.051	0.43	0.061
	Writing Task 3 (Week 7)	0.46	0.061	0.44	0.051	0.45	0.056
	Total Across Time	0.44	0.059	0.41	0.049		
Pronoun Density (Pron.Den)	Writing Task 1 (Week 3)	0.05	0.021	0.04	0.015	0.046	0.018
	Writing Task 2 (Week 5)	0.05	0.014	0.05	0.017	0.051	0.015
	Writing Task 3 (Week 7)	0.05	0.019	0.05	0.018	0.053	0.018
	Total Across Time	0.052	0.018	0.048	0.017		
All Connectives (All.Con)	Writing Task 1 (Week 3)	0.08	0.014	0.08	0.010	0.081	0.012
	Writing Task 2 (Week 5)	0.07	0.013	0.08	0.014	0.072	0.014
	Writing Task 3 (Week 7)	0.08	0.016	0.08	0.012	0.081	0.014
	Total Across Time	0.076	0.015	0.080	0.012		

N=17 students

The follow-up univariate test for time for corrected p-value (.017) showed significant effect on repeated content words only. The results for repeated content words were $F(2, 32) = 8.69$, $p = <.001$, $\eta 2 = .35$. For both groups, there were significant increases. For example, as seen in Table 4.25, the first drafts saw repeated content words increase from .41 to .46 between writing task 1 (week 3) and writing task 3 (week 7), while the last drafts increased from .39 to .44. On average, for both groups, the incidents of repeated content words increased from writing task 1 (week 3) (M=0.40) to writing task 3 (week 7) (M=0.45). No other indices were significant for the Bonferroni corrected p-value for time.

The univariate test for draft showed significant effect on repeated content words: $F(1, 16) = 15.23$, $p = .001$, $\eta 2 = .49$. For each time point, the last draft had lower repeated content words by at least .02. For example, writing task 1 (week 3), the last draft had .2 fewer repeated content words, .3 fewer for writing task 2 (week 5), and .2 fewer for writing task 3 (week 7). On average, the last drafts had lower incidents of repeated content words than the first drafts (M=0.41 and M=0.44, respectively). No other indices showed significant

effect for draft.

In summary, for changes in local cohesion across groups and time, there was a significant effect for group and time but not for group with time. While there was no significant effect for group in the follow-up analysis, there was a significant increase for adjacent sentence overlap for all words and for all nouns for both groups after the treatment. For changes in local cohesion across drafts and time for the experimental group, the analyses indicated a significant effect for draft but not for time or draft with time, but the follow-up univariate test for draft showed no significant effect. In addition, for changes in global cohesion across groups and time, there was a significant effect for time, but not for group or group with time. Specifically, there was an increase for adjacent paragraph overlap for all words and conceptual overlap of nouns between paragraphs but not for other indices for both groups after treatment. For changes in global cohesion across drafts and time for the experimental group, the analyses indicated a significant effect for time, draft, and draft with time. There was an increase in conceptual overlap of nouns between paragraphs after treatment. Although the multivariate test showed a significant effect for draft and draft with time, the univariate tests did not indicate a significant effect for draft and for draft with time. Lastly, for changes in text cohesion across groups and time, the analysis confirmed no significant interaction effect for group with time, but there was a significant effect for time and group. The follow-up analysis for group did not show any significant effect but for time indicated that all indices increased for both groups after the treatment: repeated content words, pronoun density, and all connectives. For changes in text cohesion across drafts and time for the experimental group, the analyses indicated a significant effect for time and draft but not for draft with time. Specifically, incidents of repeated content words increased after the treatment. Also, the incidents of repeated content words increased between the first and the last drafts.

4.2.6 Changes in Content

Changes in strategic measure were operationalized by rater scores for task response, which measures the quality of the response by examining if the essay presents a fully

developed position in answer to the question with relevant, fully developed and well supported ideas. A mixed ANOVA analysis was done. Mauchly's Test of Sphericity test statistic was significant for time (p = .005). Therefore, the Greenhouse Geisser correction was used. Table 4.12 summarizes the descriptive statistics for rater scores for task response by group and time.

TABLE 4.26
Descriptive Statistics for Task Response Scores by Group and Time

	Experimental		**Comparison**		**Total for Groups**	
	M	**SD**	**M**	**SD**	**M**	**SD**
Pretest (Week 1)	2.88	0.53	3.41	0.66	3.14	0.65
Writing Task 1 (Week 3)	2.81	0.36	3.25	0.66	3.03	0.57
Writing Task 2 (Week 5)	2.88	0.39	3.47	0.69	3.17	0.63
Writing Task 3 (Week 7)	3.13	0.39	3.72	0.41	3.42	0.49
Delayed Posttest (3 months after treatment)	3.34	0.44	3.88	0.34	3.61	0.47
Total Across Time	3.01	0.46	3.54	0.60		

N=16 for both comparison and experimental groups

Task response scores by groups and time: The Greenhouse Geisser corrected test statistic indicated that there was a significant effect for time: F(2.93, 120), = 10.23, p = <.001, η2 = .25. As seen in Table 4.12, for both groups, the scores for task response increased. For example, the experimental groups scores increased from 2.88 to 3.34 between the pretest and the delayed posttest while the comparison group increased from 3.41 to 3.88. On average, for both groups, the scores for task response increased from pretest (M=3.14) to the delayed posttest (M=3.61).

In addition, Greenhouse Geisser corrected test statistic indicated significant effect for group: F(1, 30), = 19.9, p = < .001, η2 = .40). For each time point, the comparison group's average score was about .4 higher. There was no significance for the interaction effect for group with time. The larger effect size for group (η2 = .40) compared to time (η2 = .25) shows that there was a large magnitude of the difference for group than across time. On average, the experimental group received lower scores for task response than did the comparison group (M=3.01 and M=3.54, respectively).

Task response scores by drafts and time for the experimental group: The initial 2 X 3 ANCOVA analysis with the number of drafts as a covariate found no main or interaction effects for the covariate. Therefore, the analysis was done again without the covariate to simplify the interpretation. Table 4.27 summarizes the descriptive statistics for rater scores for task response by draft and time.

TABLE 4.27
Descriptive Statistics for Task Response Scores by Draft and Time for the Experimental Group

	First Draft		**Last Draft**		**Total Across Drafts**	
	M	**SD**	**M**	**SD**	**M**	**SD**
Writing Task 1 (Week 3)	2.88	0.45	3.18	0.47	3.03	0.48
Writing Task 2 (Week 5)	2.91	0.40	3.35	0.55	3.13	0.53
Writing Task 3 (Week 7)	3.18	0.43	3.50	0.35	3.34	0.42
Total Across Time	2.99	0.44	3.34	0.47		

N=17 students

Mauchly"s test statistic was nonsignificant for time ($p = .75$) and interaction between time and draft ($p = .97$). The 2 X 3 ANOVA analysis indicated that there was a significant effect for time: $F(2, 32) = 7.72$, $p = .002$, $\eta 2 = .33$. As seen in Table 4.27, for each time point, the task response scores increased. For example, the scores for the first draft increased on average from 2.88 to 3.18 from writing task 1 (week 3) to writing task 3 (week 7). In addition, a similar trend can be seen for the last draft. It increased from 3.18 to 3.50 between writing task 1 (week 3) and writing task 3 (week 7). For both drafts, on average, the scores for task response increased from writing task 1 (week 3) (M=3.03) to writing task 3 (week 7) (M=3.34).

The univariate analyses also indicated that there was a significant effect for draft: $F(1, 16) = 36.13$, $p = <.001$, $\eta 2 = .69$. At each time point, the scores for task response for the last drafts were higher by at least by 0.3 points. On average, the last drafts were scored higher for task response (M=3.34 and M=2.99, respectively). There was no significant effect for interaction between drafts and time. The larger effect size for group ($\eta 2 = .69$) compared to time ($\eta 2 = .33$) shows there was a large magnitude of the difference for group than across time. There was no significant interaction effect for draft with time.

4.2.7 Changes in Quality of Language

Quality of language was examined through human ratings of grammar and lexis. To examine changes in rater scores for grammar across groups and time, a mixed ANOVA analysis was done. Mauchly's Test of Sphericity test statistic was significant for time (p = .012). Therefore, the Greenhouse Geisser correction was used. Table 4.28 summarizes the descriptive statistics for task response by group and time.

TABLE 4.28
Descriptive Statistics for Grammar Scores by Group and Time

	Experimental		**Comparison**		**Total for Groups**	
	M	**SD**	**M**	**SD**	**M**	**SD**
Pretest (Week 1)	2.97	0.29	3.47	0.50	3.22	0.47
Writing Task 1 (Week 3)	3.03	0.29	3.47	0.39	3.25	0.40
Writing Task 2 (Week 5)	3.19	0.36	3.66	0.35	3.42	0.42
Writing Task 3 (Week 7)	3.25	0.32	3.72	0.36	3.48	0.41
Delayed Posttest (3 months after treatment)	3.69	0.57	3.78	0.48	3.73	0.52
Total Across Time	3.23	0.45	3.62	0.43		

N=16 for both comparison and experimental groups

Grammar scores by group and time: The Greenhouse Geisser corrected test statistic indicated that there was a significant effect for time: F(3.04, 91.29), = 13.17, p = <.001, η2 = .305. As seen in Table 4.28, for both groups, the scores for organization increased. For example, the experimental groups scores increased from 3.13 to 3.53 between the pretest (week 1) and the delayed posttest (3 months after treatment) while the comparison group increased from 3.53 to 3.94. For both groups, on average, the scores for grammar increased from the pretest (week 1) (M=3.22) the delayed posttest (3 months after treatment) (M=3.73).

In addition, the univariate analysis indicated significance for group: F(1, 30), = 16.08, p = < .001, η2 = .349). For each time point, the comparison group's average score was higher. As seen in Table 4.10 after each time period, the difference between the comparison and the experimental group became less. For example, the difference between the scores in the pretest (week 1) was .5, but in the delayed posttest (3 months after treatment), the

difference was only .09. On average, the experimental group received lower scores from raters for task response than did the comparison group (M=3.23 and M=3.62, respectively).

There was no significance for the interaction effect for group with time: F(3.04, 91.29), = 2.19, p = .094. However, Field (2018) notes that the Greenhouse-Geisser correction is overly conservative. Therefore, the significance can be interpreted as a weak significance. Figure 4.29 shows the plots for the estimated marginal means of scores for grammar at five time points for the groups.

TABLE 4.29

Estimated Margins Means for Grammar Scores by Group and Time

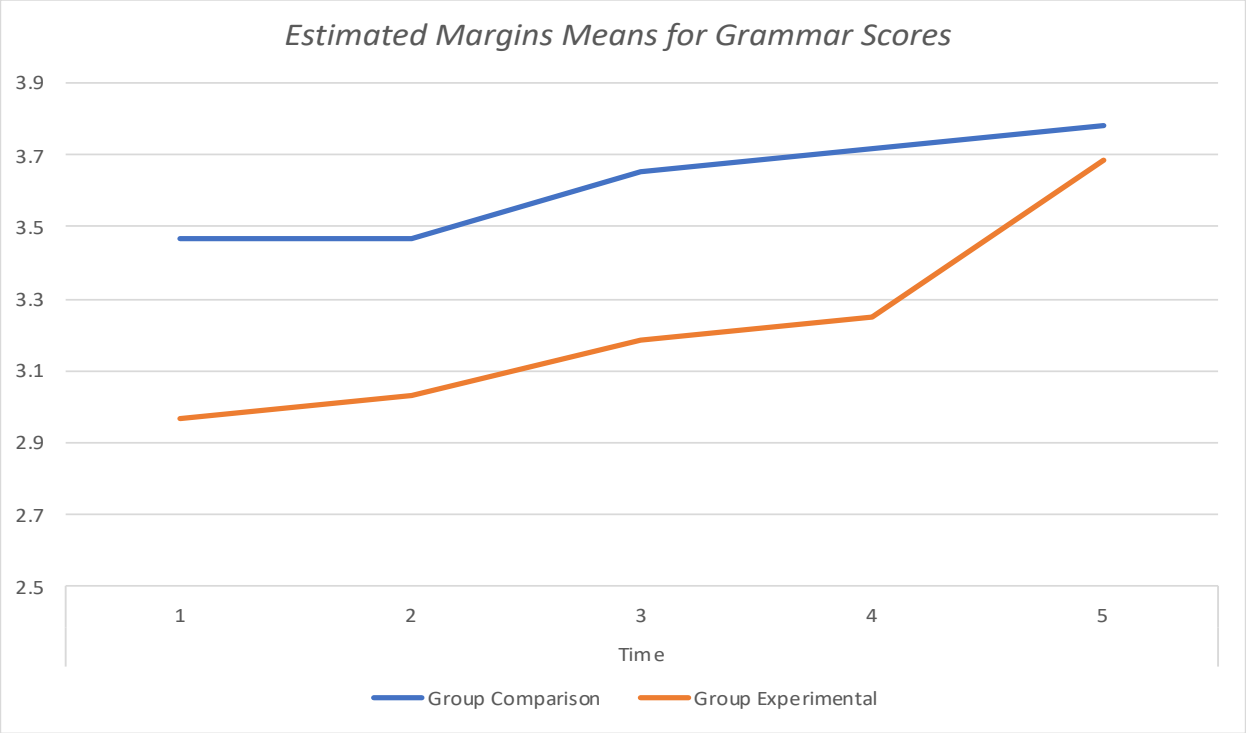


Figure 4.29 shows that while the experimental group's score for grammar started significantly lower than that of the comparison group, by the delayed posttest (3 months after treatment), the two groups" scores were almost the same, with 3.69 for the experimental group and 3.78 for the comparison group.

Grammar scores by drafts and time: The initial 2 X 3 ANCOVA test for task response showed no significance for the number of drafts as a covariate. Therefore, the analysis was done again without the covariate to simplify the interpretation. Table 4.30 summarizes the descriptive statistics for rater scores for task response by draft and time.

TABLE 4.30
Descriptive Statistics for Grammar Scores by Draft and Time for the Experimental Group

	First Draft		**Last Draft**		**Total Across Drafts**	
	M	**SD**	**M**	**SD**	**M**	**SD**
Writing Task 1 (Week 3)	3.09	0.36	3.41	0.48	3.25	0.45
Writing Task 2 (Week 5)	3.21	0.36	3.56	0.35	3.38	0.39
Writing Task 3 (Week 7)	3.27	0.31	3.50	0.35	3.38	0.35
Total Across Time	3.19	0.35	3.49	0.39		

N=17 students

Mauchly"s test statistic was nonsignificant for time (p = .14) and time with draft (p = .29). The 2 X 3 ANOVA analysis indicated that there was no significant effect for time. The univariate analysis indicated that there was a significant effect for draft: $F(1, 16) = 34.17$, $p = <.001$, $\eta 2 = .68$. At each time point, the scores for grammar for the last drafts were higher by at least 0.24. On average, the last drafts were scored higher for grammar than the first drafts (M=3.49 and M=3.19, respectively). There was no significant interaction effect for group with time.

Lexis scores by group and time: A mixed ANOVA analysis was done. Mauchly's Test of Sphericity test statistic was not significant for time (p = .41). Therefore, no correction was used. Table 4.14 summarizes the descriptive statistics for rater scores for lexis by group and time.

To examine changes in scores for lexis across group and time, a mixed ANOVA analysis was done. Mauchly's Test of Sphericity test statistic was not significant for time (p = .41). Therefore, no correction was used. Table 4.31 summarizes the descriptive statistics for rater scores for lexis by group and time.

TABLE 4.31
Descriptive Statistics for Lexis Scores by Group and Time

	Experimental		Comparison		Total for Group	
	M	**SD**	**M**	**SD**	**M**	**SD**
Pretest (Week 1)	2.81	0.31	3.56	0.51	3.19	0.56
Writing Task 1 (Week 3)	2.97	0.34	3.50	0.41	3.23	0.46
Writing Task 2 (Week 5)	3.06	0.25	3.63	0.34	3.34	0.41
Writing Task 3 (Week 7)	3.09	0.33	3.78	0.36	3.44	0.49
Delayed Posttest (3 months after treatment)	3.69	0.31	3.81	0.44	3.75	0.38
Total Across Time	3.13	0.42	3.66	0.43		

N=16 for both comparison and experimental groups

The test indicated that there was a significant effect for time: F(4, 120), = 16.32, p = <.001, η2 = .35. As seen in Table 4.31, for both groups, the scores for lexis increased. For example, the experimental group"s scores increased from 2.81 to 3.69 between the pretest (week 1) and the delayed posttest (3 months after treatment) while the comparison group increased from 3.56 to 3.81. For both groups, on average, the scores for lexis increased from the pretest (week 1) (M=3.19) to the delayed posttest (3 months after treatment) (M=3.75).

In addition, the univariate analysis indicated significance for group: F(1, 30), = 39.63, p = < .001, η2 = .57). On average, the experimental group received lower scores from raters for task response than did the comparison group (M=3.13 and M=3.66, respectively). Although univariate analysis found no significant interaction effect for group with time, for each time point, the comparison group's average score was higher, but in subsequent time points, the difference became less. Figure 4.7 shows the plots for the estimated marginal means of scores for lexis at five time points for the groups. There was no significant interaction effect for group with time.

FIGURE 4.7
Estimated Margins Means for Lexis Scores by Group and Time

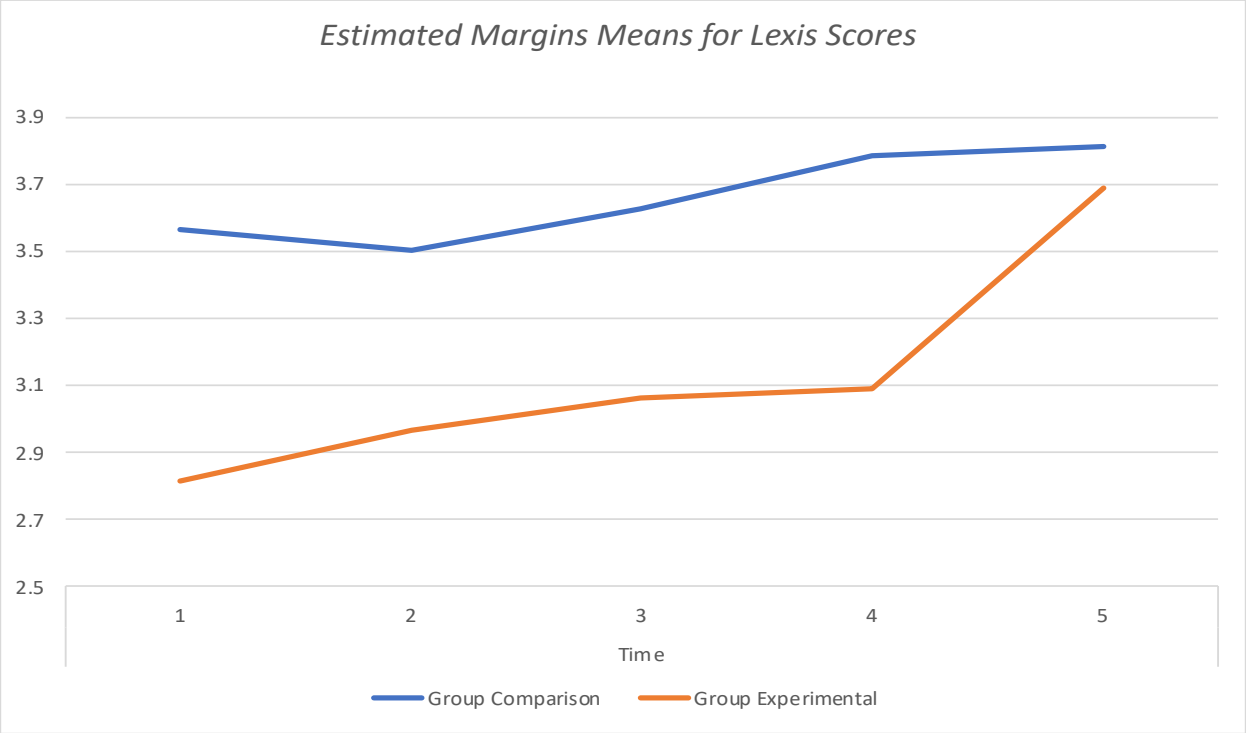


Figure 4.7 shows that while the experimental group's score for lexis started significantly lower than that of the comparison group, by the delayed posttest (3 months after treatment), the two groups' scores were almost the same, with 3.69 for the experimental and 3.81 for the comparison group.

Lexis scores by drafts and time: the initial 2 X 3 ANCOVA test for task response showed no significance for the number of revisions as a covariate. Therefore, the analysis was done again without the covariate to simplify the interpretation. Table 4.32 summarizes the descriptive statistics for rater scores for task response by draft and time.

TABLE 4.32
Descriptive Statistics for Lexis Scores by Draft and Time for the Experimental Group

	First Draft		**Last Draft**		**Total Across Drafts**	
	M	**SD**	**M**	**SD**	**M**	**SD**
Writing Task 1 (Week 3)	3.03	0.41	3.27	0.47	3.15	0.45
Writing Task 2 (Week 5)	3.09	0.26	3.35	0.42	3.22	0.37
Writing Task 3 (Week 7)	3.12	0.33	3.41	0.40	3.27	0.39
Total Across Time	3.08	0.34	3.34	0.43		

N=17 students

Mauchly"s test statistic was nonsignificant for time (p = .40) and time with draft (p = .78). The 2 X 3 ANOVA analysis indicated that there was no significant effect for time. The univariate analysis indicated that there was a significant effect for draft: $F(1, 16) = 31.02$, $p = <.001$, $\eta 2 = .66$. At each time point, the scores for lexis for the last drafts were higher by at least 0.24 points. On average, the last drafts were scored higher for lexis than the first drafts (M=3.34 and M=3.08, respectively). There was no significant interaction effect for draft with time.

4.2.8 Changes in Quality of Organization

To examine changes in scores for organization across groups and time, a mixed ANOVA analysis was done. Mauchly's Test of Sphericity test statistic was significant for time (p = .008). Therefore, the Greenhouse Geisser correction was used. Table 4.33 summarizes the descriptive statistics for rater scores for organization by group and time.

TABLE 4.33

Descriptive Statistics for Organization Scores by Group and Time both Groups

	Experimental		**Comparison**		**Total for Groups**	
	M	**SD**	**M**	**SD**	**M**	**SD**
Pretest (Week 1)	3.13	0.47	3.53	0.39	3.33	0.47
Writing Task 1 (Week 3)	3.19	0.31	3.59	0.27	3.39	0.35
Writing Task 2 (Week 5)	3.28	0.31	3.72	0.36	3.50	0.40
Writing Task 3 (Week 7)	3.25	0.32	3.78	0.36	3.52	0.43
Delayed Posttest (3 months after treatment)	3.53	0.29	3.94	0.36	3.73	0.38
Total Across Time	3.28	0.36	3.71	0.37		

N=16 for both comparison and experimental groups

Organization scores by groups and time: The Greenhouse Geisser corrected test statistic indicated that there was a significant effect across time: F(2.74, 81.72), = 10.05, p = <.001, η2 = .25. As seen in Table 4.33, for both groups, the scores for organization increased. For example, the experimental group's scores increased from 3.13 to 3.53 between the pretest (week 1) and the delayed posttest (3 months after treatment), while the comparison group increased from 3.53 to 3.94. For both groups, on average, scores for organization increased from the pretest (week 1) (M=3.33) to the delayed posttest (3 months after treatment) (M=3.73).

In addition, Greenhouse Geisser corrected test statistic indicated significance for group: F(1, 30), = 25.70, p = < .001, η2 = .46). For each time point, the comparison group's average score was about .4 higher. On average, the experimental group received lower scores from raters for organization than did the comparison group (M=3.28 and M=3.71, respectively). There was no significant interaction effect for group with time.

Organization scores by drafts and time: The initial 2 X 3 ANCOVA test for task response showed no significance for the number of revisions as a covariate. Therefore, the analysis was done again without the covariate to simplify the interpretation. Table 4.34 summarizes the descriptive statistics for rater scores for task response by draft and time.

TABLE 4.34

Descriptive Statistics for Organization Scores by Draft and Time the Experimental Group

	First Draft		Last Draft		Total Across Drafts	
	M	**SD**	**M**	**SD**	**M**	**SD**
Writing Task 1 (Week 3)	3.27	0.44	3.32	0.39	3.29	0.41
Writing Task 2 (Week 5)	3.29	0.31	3.50	0.31	3.40	0.32
Writing Task 3 (Week 7)	3.27	0.31	3.65	0.23	3.46	0.33
Total Across Time	3.28	0.35	3.49	0.34		

N=17 students

Mauchly"s test statistic was nonsignificant for time (p = .1) and time with draft (p = .74). The 2 X 3 ANOVA analysis indicated that there was no significant effect for time. The univariate analysis indicated that there was a significant effect for draft: $F(1, 16) = 21.16$, $p = <.001$, $\eta 2 = .57$. At each time point, the scores for grammar for the last drafts were higher. In writing task 1 (week 3), the difference was 0.059; in writing task 2 (week 5), the difference was .21; and in writing task 3 (week 7), the difference was .38. On average, the last drafts were scored higher for organization than the first drafts (M=3.49, SD=0.35 and M=3.28, SD=0.35, respectively). There was no significant interaction effect for draft with time.

4.2.9 Revisions across Drafts and Tasks

Table 4.35 reports descriptive statistics for the number of drafts submitted and the results from the ratings of the magnitude and effects of revisions across drafts for the experimental group. The rating scale included two items: the magnitude of revisions in the first and the last drafts submitted (0=no change, 1= minimal changes, and 2=substantive changes) and the effects of revisions on writing quality (0=no effect, 1=negligible effect, 2=mixed effect, and 3=positive effect). the 17 students in the experimental group, on average, submitted 4.25 drafts for all the writing tasks. As seen in Table 4.35, the average number of drafts submitted differed by writing task. The greatest number of drafts submitted were made for Writing Task 2 at 5.12 and the fewest for Writing Task 3 at 3.71.

TABLE 4.35
Descriptive Statistics for Number of Drafts, Change and Effect of Revisions

	Number of Drafts Submitted		**Magnitude of Changes Between the First and Last Drafts**		**Effect of Changes on Writing Quality**	
	M	**SD**	**M**	**SD**	**M**	**SD**
Writing Task 1 (Week 3)	3.94	2.61	1.35	0.61	2.24	0.90
Writing Task 2 (Week 5)	5.12	2.01	1.53	0.51	2.65	0.61
Writing Task 3 (Week 7)	3.71	2.05	1.53	0.51	2.47	0.62
Total	4.25	2.31	1.47	0.54	2.45	0.73

N=17 students

Furthermore, the scores for the magnitude of revisions between the first and the last drafts stayed the same between writing task 2 (week 5) and writing task 3 (week 7) while the average number of drafts submitted decreased by 1.41 indicates that students made more substantive revisions than minimal revisions in later writing tasks. In general, the students made more frequent and extensive changes between the first and the last drafts. In writing task 1 (week 3), the rating scale for magnitude of changes between the first and the last drafts shows that the rating was 1.35 (SD=.61) while in writing task 3 (week 7), it was 1.53 (SD=.51) showing an increase. In addition, these changes had a positive effect on writing quality. For writing task 1 (week 3), which had the lowest rating for magnitude of changes between the first and the last drafts, the rating for effect of changes on writing quality was 2.24 (SD=.90). The rating for the effect of changes on writing quality for writing task 2 (week 5) and writing task 3 (week 7) are higher indicating that the more changes students made, the higher writing quality. In other words, the changes the students made to their writing had a positive effect on writing quality, and the more substantive changes the students made between the first and the last drafts, the greater the positive effect on writing quality.

4.2.10 Changes in Criterion Trait Scores

Changes in Criterion scores could not be analyzed because they lacked variability. Criterion assigns each essay one of three levels (developing, proficient, or advanced) on

each of three criteria 1) word choice, 2) Conventions (Grammar, Usage, and Mechanics), and 3) organization, development, and style.

For the comparison class, there were a total of 80 scores. Each participant (n=16) wrote five essays: pretest, test 1, test 2, test 3, and delayed posttest. For the experimental class, there were a total of 135 scores. Each participant (n=17) received eight scores: pretest, task 1 first draft, task 1 last draft, task 2 first draft, task 2 last draft, task 3 first draft, task 3 last draft, and delayed posttest. One participant was missing a delayed posttest. Table 4.36 reports the distribution of scores for Criterion trait scores for both groups. As Table 4.36 shows, the great majority of the essays received a rating of proficient from Criterion on each of the three criteria.

Changes in Criterion scores could not be analyzed because they lacked variability. Criterion assigns each essay one of three levels (developing, proficient, or advanced) on each of three criteria 1) word choice, 2) Conventions (Grammar, Usage, and Mechanics), and 3) organization, development, and style.

TABLE 4.36
Distribution of Criterion Trait Scores

Comparison Group			
	Criterion Score for Word Choice	**Criterion Score for Grammar, Usage, and Mechanics - Conventions**	**Criterion Score for Organization, Development, and Style**
Developing	0	1	0
Proficient	79	73	80
Advanced	1	6	0
Experimental Group			
Developing	0	2	0
Proficient	128	122	135
Advanced	7	11	0

N=16 for the comparison and N=17 for the experimental groups

4.3 Students' Views of Criterion and Hybrid Feedback

Overall findings from the questionnaire and the focus-group interview indicated that the students in the experimental group generally held positive attitudes toward using AWE to improve their writing. In general, students found Criterion feedback helpful but found the combination of Criterion and teacher feedback to be more specific and contingent to students' needs and proficiency levels.

The Perception of Criterion Questionnaire elicited responses for three different dimensions: usefulness of scoring, the usefulness of feedback, and overall perception of Criterion. The results indicated that students in the experimental group found both the holistic and trait scores unhelpful in evaluating their performance but found most feedback helpful in revising their essays. Overall, the students were satisfied with using Criterion to improve their writing. Table 4.37 shows the mean and standard deviation for the responses for each question in the questionnaire. It shows that the means for the first three questions, which deal with Criterion scoring, are low (<2.53 out of 5), indicating students" dissatisfaction. Other responses were all over 3 with the exception of Q9 (I found the Criterion feedback on Style helpful), which scored only 2.76. Four questions received scores of over 4: Q4 (I found using Criterion feedback helpful in revising my essays), Q5 (I found using Criterion feedback helpful in revising my essays), Q8 (I found the Criterion feedback on Mechanics helpful), and Q13 (I think I will use Criterion again in the future if I have the chance).

TABLE 4.37

Descriptive Statistics of Experimental Group's Perception of Criterion

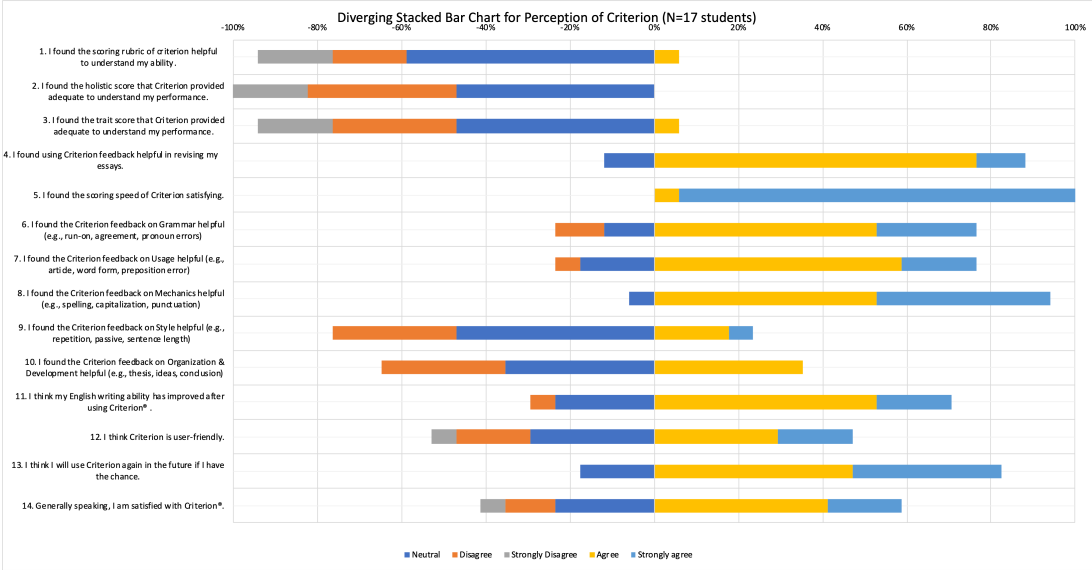
	Mean	**SD**
1. I found the scoring rubric of Criterion helpful to understand my ability.	2.53	0.87
2. I found the holistic score that Criterion provided adequate to understand my performance.	2.29	0.77
3. I found the trait score that Criterion provided adequate to understand my performance.	2.41	0.87
4. I found using Criterion feedback helpful in revising my essays.	4.00	0.50
5. I found the scoring speed of Criterion satisfying.	4.94	0.24
6. I found the Criterion feedback on Grammar helpful (e.g., run-on, agreement, pronoun errors).	3.88	0.93
7. I found the Criterion feedback on Usage helpful (e.g., article, word form, preposition error).	3.88	0.78
8. I found the Criterion feedback on Mechanics helpful (e.g., spelling, capitalization, punctuation).	4.35	0.61
9. I found the Criterion feedback on Style helpful (e.g., repetition, passive, sentence length).	2.76	0.83
10. I found the Criterion feedback on Organization & Development helpful (e.g., thesis, ideas, conclusion).	3.06	0.83
11. I think my English writing ability has improved after using Criterion.	3.82	0.81
12. I think Criterion is user-friendly.	3.35	1.17
13. I think I will use Criterion again in the future if I have the chance.	4.18	0.73
14. Generally speaking, I am satisfied with Criterion.	3.53	1.12

N=17 students. Strongly disagree = 1, disagree = 2, neutral = 3, agree = 4 and strongly agree = 5

Figure 4.8, a diverging stacked bar, shows students" responses to the three dimensions more clearly. The first three questions (questions 1-3) deal with the usefulness of scores and the scoring rubric. The students were either neutral or dissatisfied with AWE scores and the rubric. The mean score for the three items is 2.41 out of 5. Students felt neutral (M=2.53, SD=0.87) about the rubric and found the holistic scores that Criterion provided unhelpful (M=2.29, SD=0.77). The trait scores faired marginally better (M=2.41, SD=0.87).

FIGURE 4.8

Diverging Stacked Bar Chart for Items on the Perception of Criterion Questionnaire



N=17 students

Second, the next 7 questions (questions 4-10) deal with the usefulness of Criterion feedback. The mean score for the seven items is 3.84 indicating overall satisfaction with the usefulness of Criterion feedback. For feedback on grammar, usage, and mechanics, the mean score for the three items (questions 6-8) is 4.04. However, the responses for feedback on style and organization were more muted, with the mean score for the two items (questions 9-10) being 2.91. As expected, the students found the scoring of Criterion satisfying (M=4.94, SD=0.24). In addition, students found Criterion helpful in the revision process (M=4.00, SD=0.50).

Lastly, the last 4 questions (questions 11-14) deal with students' overall perception of Criterion. The mean score for the four items is 3.72 indicating overall satisfaction with Criterion. For general satisfaction with Criterion, students, on average, gave a scores of (M=3.53, SD=1.12). When asked to identify their level of agreement with the following statement, "I think my English writing ability has improved after using Criterion", on average, students gave a high score (M=3.82, SD=0.81). Although students strongly agreed with using Criterion in the future if they have the chance (M=4.18, SD=0.73), there was less agreement about the user-friendliness of Criterion (M=3.35, SD=1.17). The

high SD for user friendliness of Criterion shows that there were a lot of diverging opinions.

During the focus-group interview four themes emerged about students" perception of hybrid feedback: affordances of Criterion feedback, impact on the revision and writing process, teacher feedback vs Criterion feedback, and constraining factors to using Criterion. Each of these themes is discussed in the following paragraphs.

The first theme is affordances of Criterion feedback. For feedback types, students generally gave comments on the types of feedback that Criterion gave on mechanics, organization & development, style, and grammar. All students found surface-level feedback on punctuation and spelling useful, and most reported that they already use similar features in word and Grammarly. For organization, most students commented that Criterion highlighted introductory material, thesis statement, main ideas, supporting ideas, conclusion, and transitional words and phrases, which helped them think more about supporting ideas and linking sentences; also, just over half of the students found generic feedback to be useful for reflection. For style, almost all students found the feedback on the repetition of words and most found passive voice to be frustrating. Specifically, Criterion highlighted passive voice, but the students did not understand from the feedback why passive voice may not be appropriate because of the generalized feedback, "You have used the passive voice in this sentence. Depending upon what you wish to emphasize in the sentence, you may want to revise it using the active voice." For grammar, students highlighted the fact that Criterion did not catch all errors; upon revising, they caught errors that were not caught by Criterion. However, students were used to such occasions because many of them use Grammarly and other grammar checkers. Some students were proactive, and when they could not resolve the errors, they went to external sources of information, such as their friends. However, most students noted that the more they used the tool and became familiar with it, they were able to fix more mistakes.

Interestingly, the students mentioned that the game-like mechanics of finding the errors motivated them to fix more errors because it was "like a game." However, more students were motivated to fix all mistakes because they wanted the essay to be better than previous versions. Overall, a majority of students found that the integration of Criterion in the classroom helped to understand their weaknesses in their writing.

A second theme that emerged from the focus group is the impact of hybrid feedback on students" revision and writing processes. More than half of the students responded that the combination of AWE and teacher feedback helped them focus on the audience and think about what questions the reader may have for their writing. After feedback from the teacher, more than half of the students focused on the requirements of the task to answer the essay prompt, thought more about the type of the essay they are writing, and, overall, conceptualized how their ideas could be developed. The majority of students iterated that the Criterion feedback highlighting the organizational aspect of writing was useful because it helped them think more about ideas and development. For organization, the integration of feedback helped students see that the pre-writing phase of brainstorming is an integral part of the writing process and focused their attention on the essay's overall coherence. For monitoring and revising, having opportunities to revise their essays helped the students see improvement between the first and the last drafts.

The students reported that, in previous classes, the class structure did not allow time for revision, and that they only experienced writing assessments that emphasized writing as a product. They reported that the higher requirement for length in the class and the combination of feedback helped them appreciate the importance of brainstorming to produce a more coherent essay. The teacher feedback and highlighting organizational aspects by Criterion helped students answer the essay prompt more directly and think more about developing and supporting their ideas reflecting assessment as/for learning. In addition, the Criterion feedback on grammar and lexis helped the students to think more about the appropriacy of the lexis and syntactic structures they use. However, they found it frustrating not to have direct feedback on word choice.

Nonetheless, for some errors such as ending punctuation and spelling, the students reported that they did not internalize the feedback because these were mistakes that they believed they could fix independently or because word processors already provide such feedback. Overall, all students agreed that the opportunities for revision helped them understand that their work is "never finished," and that revision is part of the writing process.

The third theme is comparison of teacher and Criterion feedback. Almost all students enjoyed the immediacy and convenience of feedback from Criterion because they were able to make revisions and corrections as both topic and writing were still fresh in their minds and found it to be motivating because some mistakes could be found and resolved immediately. Most students agreed that the number of highlighted errors for each category gave them an idea of their weaknesses. All students agreed that they did not rely on trait or overall scores because there was no variability of the scores, so they relied more on feedback. Some students mentioned that even after taking the assignment home and utilizing other resources, they still needed direct, specific, and personalized feedback from the teacher to resolve the errors because that was easier to understand and was more specific. Accordingly, most students agreed that teacher feedback was needed for students with less proficient grammar to clarify Criterion feedback.

Students found teacher feedback to be more specific than Criterion feedback and to be more clearly delineated to help students improve specific problem areas. Also, students focused on the teacher's scores because the marks would ultimately impact their overall grades. Most students agreed that teacher feedback was more useful than time-on-task for developing ideas and that highlights from Criterion helped them to focus on structure. Students noted that they focused on teacher feedback more because they "understand better because the words used in the class is the same as the feedback," they can ask questions and "feedback is easier to understand," and "teacher gives feedback in [sic] most serious issues and not all." All students agreed that grammar feedback was more

useful from the teacher, but some students mentioned that they were too nervous to ask questions in teacher conferences.

The students reported that , in previous classes, teachers' feedback practices were wide-ranging from direct, to indirect, from a paragraph summarizing problems, to a simple "good job" with a score. The feedback timeline ranged from a week to a month. For example, some teachers in previous writing courses gave very little feedback, while some wrote a couple of sentences at the end of the essay. However, the students accepted little feedback because they understand that it is a time-consuming process. Also, some teacher feedback was difficult to resolve because it was not highlighted or linked to the text as they were in Criterion. When asked if Criterion could replace teacher feedback, all students replied that it could not replace teacher feedback but believed that it could replace teacher feedback that only had scores or a single sentence feedback. Most students suggested all teacher feedback could be more systematic and give clear advice for improvement. When asked if they would prefer teacher feedback or combined teacher and Criterion feedback, all students would prefer the hybrid form; most students felt that the combination helped them to be more independent and motivated them to revise. Overall, all students agreed that the hybrid feedback changed their orientation for writing as a process rather than a product: students agreed with one student's description of previous writing practice - "Just write. Don't think. Don't care."

The fourth theme concerns factor that students felt could constrain the use of Criterion. The four most frequently reported constraining factors were 1) usability of Criterion, 2) problems with feedback, 3) problems with scoring, and 4) individual learner factors. For usability, most students found the onboarding process of using the app confusing due to the UX not being friendly. The multiple windows and tabs caused confusion, and students found the overall presentation "too busy." In addition, some students would have preferred more hyperlinks to external resources that would facilitate learning. For feedback problems, the students found the feedback to be too general or that they could not understand it for it to be useful because unlike the teacher, Criterion does not provide

models, alternatives, or examples suited for the students' proficiency.

Almost all the students reported instances when they did not understand the feedback. Even when they understood the feedback, they could not resolve it because there were no remediation tools. In addition, they found the feedback to be too general and found the lack of variance to be troubling: Criterion highlighted a problem area but did not give specific guidance for improvement; also, they found the feedback too similar after each revision, and Criterion did not give specific feedback or how a new version of the revision was improved compared to the previous version. The most significant frustration seems to have stemmed from Criterion not giving systematic advice on how the new revision is better than the previous version. Criterion would highlight problem areas but did not offer specific guidance for improvement. However, it offered generic comments and referred them to read the writer"s handbook or ask a teacher for guidance. For example, for a prompt on liberal arts education, Criterion highlighted the word "liberal" and gave the following feedback, "You have repeated these words several times in your essay. Your essay will be stronger if you vary your word choice and substitute some other words instead. Ask your teacher for advice." The students did not know any synonyms for liberal arts and would have preferred an integration of a thesaurus. Likewise, even when students understood the feedback, they did not know how to resolve them.

Almost all students found the scoring aspect frustrating because the overall scores were too general to interpret and were confused when interpreting trait scores because there was no variance in scores or the corresponding feedback. For example, a few stated that they could not raise their trait scores even after multiple attempts at improving their essays by resolving every highlighted point. Moreover, the students did not understand the differences between overall scores and did not understand why an essay was given a five rather than a six because the scoring descriptions were too general. Lastly, some students mentioned lack of time to use Criterion due to homework for the listening, reading, and writing components of the course or external commitments such as jobs. Others reported that they did not trust Criterion feedback because they do not believe that

machines could understand their writing.

4.4 Selected Individual Case Analyses

In this section, I present an in-depth analysis of data from three students from the experimental group. The three cases selected are (1) Ben, who made substantive revisions but did not engage with AWE feedback and relied more on his grammatical knowledge; (2) Rebecca, who reported that she engaged more with machine feedback due to its convenience of being able to receive feedback when and where she wanted; and (3) Jasmine, who consistently engaged fully with both teacher and AWE feedback and strove to resolve all errors. Table 4.38 summarizes the overall revision behaviour and attitudes towards teacher and AWE feedback of the three students.

TABLE 4.38
Description of Overall Revision Behaviour and Attitudes Towards Teacher and AWE Feedback.

	Average Number of Drafts Submitted	**Magnitude of Change in Drafts**	**Average Increase of Combined Rater Scores Between First and Last Drafts**	**Attitude Towards Teacher Feedback**	**Attitude Towards AWE Feedback**
Rebecca	6 (4 min, 8 max)	Both surface-level and changes in content and organization	.17	Neutral	Positive
Ben	2.67 (1 min, 4 max)	Both surface-level and changes in content and organization	.17	Positive	negative
Jasmin	8.67 (7 min, 10 max)	Mostly surface-level changes	.29	Positive	Positive

In addition, the demographic profiles of the three students are summarized in Table 4.39.

TABLE 4.39
Demographic Profiles of the Three Students

	Rebecca	**Ben**	**Julia**
Country	Mainland China	Mainland China	Mainland China
Language	Mandarin	Mandarin	Mandarin
Gender	Female	Male	Female
Age	18	20	20
Overall IELTS Score	5.5	NA	6
Writing IELTS Score	5.5	NA	6
Years studying English	13	9	10
Length of stay in Canada in Weeks	9	9	17
Chosen major at university	Digital Media	Kinesiology	Finance

At the start of the program, a questionnaire on students' attitudes towards teacher feedback and computer feedback, if any, was collected. Table 4.40 reports the results for three focal students. The responses are rated on a four-point Likert scale (Definitely agree = 4, Mostly agree = 3, mostly disagree = 2, definitely disagree = 1).

TABLE 4.40
Results for questionnaire on Student Attitudes Towards Teacher and Computer Feedback

	Rebecca	**Ben**	**Julia**
I think doing more writing is important to improve my writing.	2	4	4
I pay attention to the score when my writing is returned.	4	3	4
I pay attention to the feedback when my writing is returned.	3	4	4
I think the feedback I received from my instructors was timely.	2	2	2
I try to avoid similar problems in future writing when I receive feedback.	4	4	4
I revise my essays before submission.	2	3	4
I think revising my essays is an important part of the writing process.	2	3	4
I like revising my essays.	2	3	4
I find instructor feedback helpful when revising my essays.	2	4	4
I find peer feedback helpful in revising my essays.	2	3	3
I have previous experience with computer feedback systems (e.g. Grammarly, Microsoft word grammar, spelling checked, turnitin.com, etc…)	3	3	4
If yes to the previous question, I find computer feedback helpful in revising my essays.	4	2	4

Case 1: Rebecca

Class observations suggested that Rebecca was a diligent student but did not initiate

questions and was shy about responding to teacher-initiated questions. Although she was responsive to teacher feedback during teacher conferences, she did not ask questions for elaborations and clarifications. However, her enthusiasm for learning was evident, as seen by the number of drafts she submitted.

Rebecca reported that she has been in Canada for nine weeks before the program start, and she has been accepted to the Digital Media Program at the university. Rebecca entered the program with an overall IELTS score of 5.5, and her writing score was also 5.5.

As seen from Table 4.40, Rebecca did not think that writing practice was essential for her writing improvement or that the writing process is necessary. She also does not like to revise her essays. She also did not find teacher feedback as helpful as the other two students but found computer feedback very helpful in revising essays before the treatment. Although not in the curriculum, Rebecca reported finding peer feedback unhelpful. Notwithstanding the fact that she did not find the feedback useful, she paid attention to it and tried to avoid similar problems in the future. However, she reported paying more attention to her scores than feedback as motivation to improve her writing. In addition, she mostly disagreed that revising her essays is an integral part of the writing process.

For revisions, Rebecca submitted the second-highest number of drafts in the experimental group. She submitted an average of six drafts per assignment: eight for writing task 1, six for writing task 2, and four for writing task 3. The changes in her revisions were all substantive, meaning that she made both high and low-level revisions. The effects of the revisions that she made were mixed to positive. Table 4.41 reports the number, magnitude of change, and effect of revisions between the first and the last drafts on writing quality for three writing tasks.

TABLE 4.41
Number, Magnitude of Change, and Effect of Revision between First and Last Drafts on Writing Quality for Rebecca

	Writing Task 1	**Writing Task 2**	**Writing Task 3**
Number of Revisions	8	6	4
Magnitude of Change in Revision	Substantive	Substantive	Substantive
Effect of Revision	Mixed	Mixed	Positive

During her revisions, Rebecca strived to resolve all errors identified by Criterion by reflecting on the feedback, consulting external sources, and asking about specific grammar points in class; on average, she resolved 86 percent of all errors in her last submissions from the first. Table 4.42 summarizes the number of mechanical, grammar, usage, and total errors by writing tasks for the first and the last drafts.

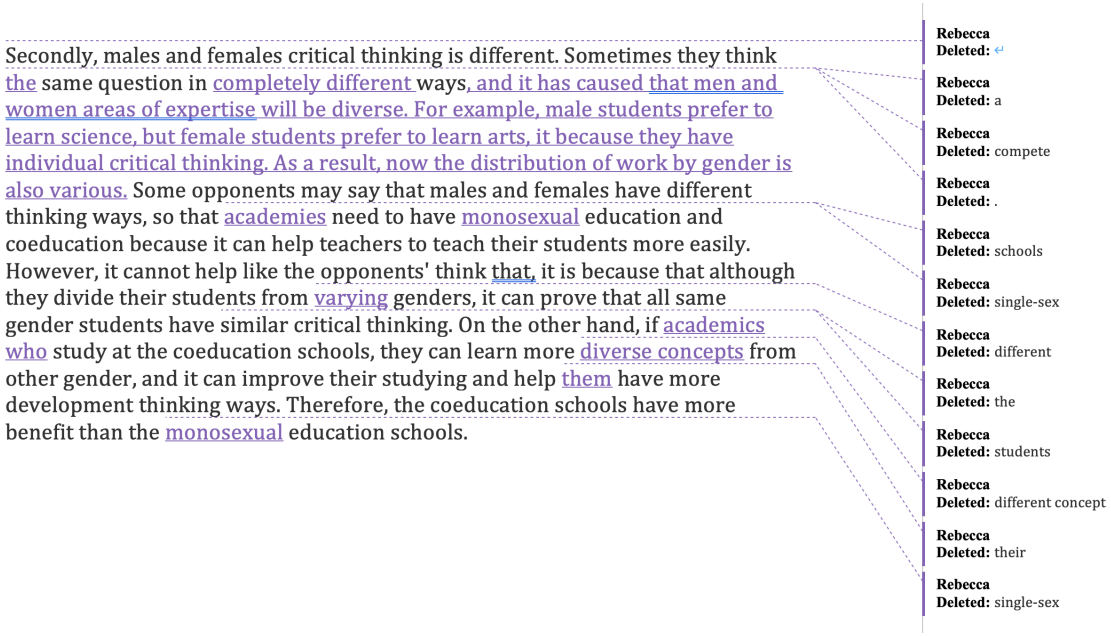
TABLE 4.42
Number of Errors in First and Last Drafts in Mechanics, Grammar, and Usage from Criterion for Rebecca

	Writing Task 1		**Writing Task 2**		**Writing Task 3**	
Revision	First	Last	First	Last	First	Last
Mechanics	5	2	3	0	2	6
Grammar	2	0	3	0	2	0
Usage	18	0	20	0	15	2
Total	25	2	26	0	19	8

The analysis of Rebecca"s revisions showed that she made changes to surface level concerns, content, and development. During the interview, she suggested that she incorporated teacher feedback because, ultimately, the scores would be given by the teacher. In addition, she revealed that she believes that the "teacher gives feedback in most serious issues and not all." She reported that this belief led her to incorporate both types of feedback equally. A side-by-side comparison of her writing showed that this was the case. She incorporated teacher feedback from previous writing tasks in later tasks. For example, as seen in Figure 4.9, which shows the changes between the first and the last drafts for writing task 3, Rebecca added an example following teacher feedback from previous assignments suggesting that she help readers understand her ideas more clearly

by giving examples. The other changes in the examples were for Criterion feedback regarding word choice: Criterion identified the repetition of the phrase 'single-sex'.

FIGURE 4.9
Sample of High- and Low-Level Revisions for Rebecca



Rebecca"s scores correspondingly increased between the first and the last drafts. Specifically, her scores increased most consistently for lexis and grammar. For writing task 1, her score for lexis increased, and for writing task 3, her scores for lexis and grammar increased. Table 4.43 summarizes Rebecca's scores for the first and the last drafts for task response, organization, lexis and grammar.

TABLE 4.43
Rebeccas's Scores for First and Last Drafts for Task Response, Organization, Lexis, and Grammar

	Writing Task 1		**Writing Task 2**		**Writing Task 3**	
	First	Last	First	Last	First	Last
Task Response	2.5	2.5	3	3.5	3.5	3.5
Organization	3	3	3.5	3.5	3.5	3.5
Lexis	3	3.5	3	3	3	3.5
Grammar	3	3	3	3	3	3.5

In new pieces of writings, after a significant increase of errors after the pretest, the number of errors found by Criterion steadily decreased for all three types of errors in general. Table 4.44 reports the number of errors in new writings in mechanics, grammar, and usage from Criterion.

TABLE 4.44
Number of Errors in New Writings in Mechanics, Grammar, and Usage from Criterion for Rebecca

	Mechanics	**Grammar**	**Usage**	**Total**
Pretest	0	6	5	11
Writing Task 1	5	2	18	25
Writing Task 2	3	3	20	26
Writing Task 3	2	2	15	19
Delayed posttest	2	1	3	6

Rebecca reported that she made a concerted effort to address both teacher and machine feedback by attending to Criterion feedback on mechanics, grammar and usage and focusing on teacher feedback on content and organization. However, although her scores for task response and organization showed improvement during the treatment, her scores for lexis and grammar did not show any improvement until the delayed posttest. Table 4.45 reports Rebecca"s scores for new writings on task response, organization, lexis, and grammar. It shows that her scores tended to stay the same or increase over time.

TABLE 4.45
Rater Scores for New Writings in Task Response, Organization, Lexis, and Grammar for Rebecca

	Task Response	Organization	Lexis	Grammar
Pretest	2.5	3	3	3
Writing Task 1	2.5	3	3	3
Writing Task 2	3	3.5	3	3
Writing Task 3	3.5	3.5	3	3
Delayed posttest	3.5	3.5	4	4

The results of the Writing Process Questionnaire indicate that Rebecca reported having improved marginally in terms of two phases: conceptualization and monitoring and revising at low-level. Her scores for both changed less than .7. However, other phases changed significantly, especially for generating texts phase (increase of 1.6) and revising at high-level phase (increase of 1.3) indicating that she thought more about the task, grammar, and organization before she started writing and during the revision process after instruction.

During the interview and in her answers to the Perception of Criterion Questionnaire, Rebecca suggested that she, in general, preferred machine feedback because it integrates better with her workflow due to the instantaneous nature of feedback. She also preferred the comprehensive nature of machine feedback. She stated that "I think a teacher gives feedback in most serious issues and not all." However, she felt that there was a disconnect between teacher and Criterion feedback:

> [When] I received a good score from Criterion but when I received your feedback, you pointed out that there are problems with the meanings of the words, and how some words were too general.

In addition, Rebecca found generic feedback frustrating because she did not know how to respond to it. She especially found Criterion highlighting repetition of words unhelpful. In general, she found the low-level feedback from Criterion about grammar, usage, mechanics very helpful and would continue to use such feedback in the future.

During the delayed posttest, Rebecca mentioned that she actively sought out other automated feedback and correction systems to help her write her assignments for university studies. She employed Grammarly, a cloud-based writing assistant that reviews spelling, grammar, and punctuation, routinely for her writing assignments and a Chinese AWE feedback system. Rebecca reported not seeking out any writing help resources such as the ESL open learning centers, workshops, or the writing center for support during her university semester.

Overall, Rebecca"s view of hybrid AWE feedback was more positive than receiving instructor feedback only. In the interview and the Writing Process Questionnaire, she reported that the hybrid feedback helped her improve her writing. Specifically, she reported that the combination of feedback helped prioritize her ideas, and she focused more on organization and development. Rebeca also responded very favourably to the Perception of Criterion Questionnaire; she strongly agreed that Criterion was helpful in her writing development, with the exception of its scoring. The analysis of her written work shows that she revised both surface-level and content and organization, and the revisions were all substantive, incorporating both teacher and machine feedback. She strived to correct the errors that Criterion highlighted between her drafts and in new pieces of writing; she steadily made fewer mistakes in mechanics, grammar, and usage errors. In new pieces of writing, compared to the pretest, in the delayed posttest, she connected her ideas more logically, developed her ideas more fully, made fewer grammatical mistakes, and was more cognizant of her word choice.

Case 2: Ben

Class observations suggested that Ben was a diligent student but had issues with time management. He was the only one in the class with a part-time job. He worked an average of 20 to 30 hours a week. However, he submitted all his assignments on time and made substantive changes except for the last writing task. During teacher conferences, he was dedicated to improving his language ability because, unlike other students who

would be returning to China after graduation, he wanted to immigrate to Canada, and his English ability would be a crucial component for his future plans. He asked questions during conferences and asked for clarifications on grammatical rules.

Ben reported that he had been in Canada for nine weeks before the program started and has been accepted to the Kinesiology Program at the university. Ben did not enter the program with an IELTS score.

As seen from Table 4.40, Ben found teacher feedback important to his writing development but computer feedback to be unhelpful. Ben also reported that instructor feedback was helpful for revising his essays and that he does revise his essays before submitting them. However, he found his instructor feedback was not timely in the past and mostly agreed that revising essays is an important part of the writing process.

For revisions, Ben submitted on average two and a half drafts. He submitted one for writing task 1, four for writing task 2, and three for writing task 2. The changes in his revisions were all substantive, meaning that he made both high and low-level revisions. However, after examining his drafts, Ben seemed to have disregarded most of Criterion feedback in writing tasks 1 and 2. For writing task 3, he made minimal changes and seemed to have accepted Criterion feedback. In the interview, he stated that for writing task 3, he did not have time to make substantive changes due to increased commitments to his part-time job. Surprisingly, the substantive changes he made in writing tasks 1 and 2 had a mixed effect on the overall quality of his writing, while the minimal modifications he made in writing task 3 had a positive impact. Table 4.46 reports the number, magnitude of change, and effect of revisions between the first and the last drafts on writing quality for three writing tasks.

TABLE 4.46
Number, Magnitude of Change, and Effect of Revision between First and Last Drafts on Writing Quality for Ben

	Writing Task 1	**Writing Task 2**	**Writing Task 3**
Number of Revisions	1	4	3
Magnitude of Change in Revision	Substantive	Substantive	Minimal
Effect of Revision	Mixed	Mixed	Positive

During his revisions, on average, Ben resolved 58 percent of all errors identified by Criterion between his first and last submissions. Table 4.47 summarizes the descriptive statistics for number of revisions, number of mechanical, grammar, usage, and total errors by writing task.

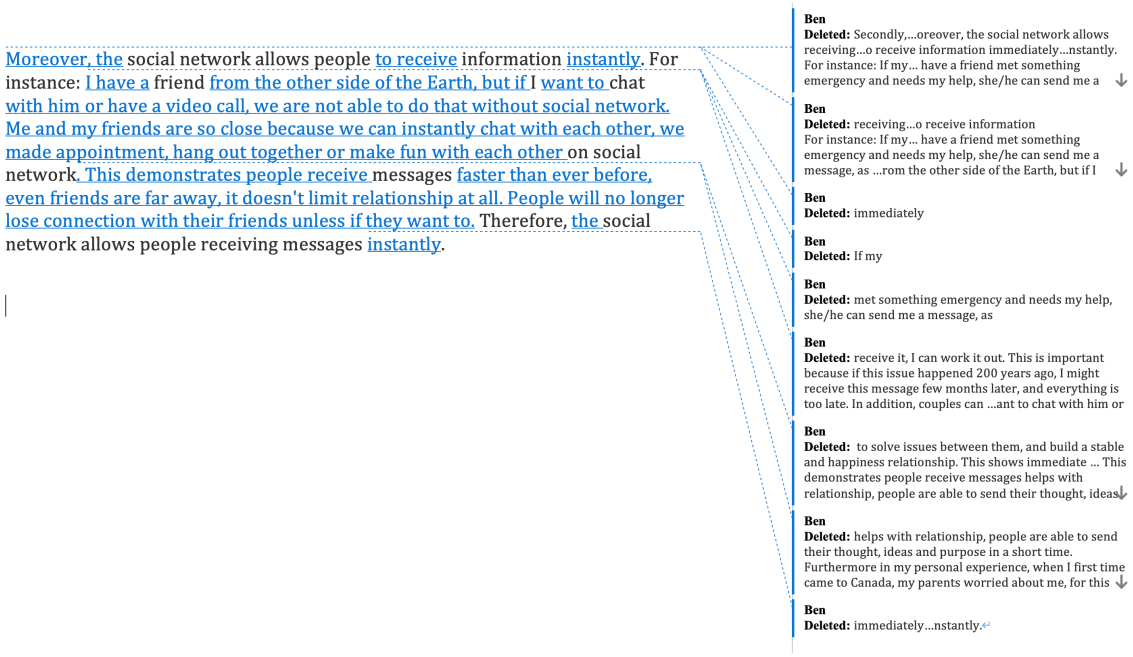
TABLE 4.47
Number of Errors in First and Last Drafts in Mechanics, Grammar, and Usage from Criterion for Ben

	Writing Task 1		**Writing Task 2**		**Writing Task 3**	
Revision	First	Last	First	Last	First	Last
Mechanics	7	1	8	2	4	6
Grammar	6	2	2	1	5	0
Usage	17	7	5	11	10	7
Total	30	10	15	14	19	13

The analysis of Ben"s revisions showed that he frequently disregarded corrective feedback generated by Criterion in writing tasks 1 and 2. When prompted why, he explained that "machines cannot correct your ideas." He didn't trust the AWE system to give correct feedback, so he often chose to look for errors independently. In writing task 2, he only made surface-level changes because he was "too busy to look" for errors due to other commitments. For instance, in writing task 1, he significantly revised his essay. As can be seen in Figure 4.10, which shows the changes from his first to the last draft, he mostly rewrote the paragraph making the topic development more succinct and removing repetition of ideas. He also found errors that were not identified by Criterion and resolved them. Here, he fixed the mistake in gerund that was not part of Criterion feedback: "Secondly, social network allows people receiving information immediately" to

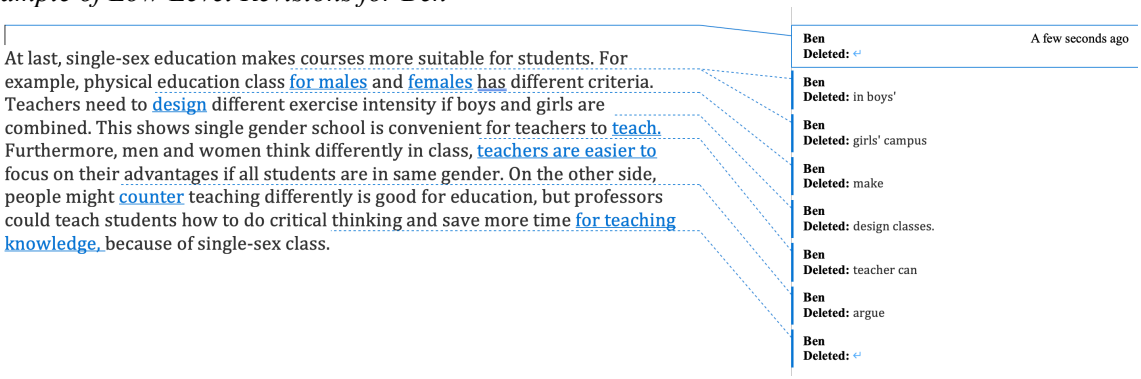
"Moreover, the social network allows people to receive information instantly." However, he introduced new errors, such as the misuse of articles.

FIGURE 4.10
Sample of Writing High-Level Revisions for Ben



Ben did not have time to revise the third writing task as much as he would have liked to. The revisions he made mainly were related to word choice issues that Criterion identified. He edited them to be more academic and less repetitive: he changed boys to males and argue to counter as shown in Figure 4.11, which is a paragraph from writing task 3.

FIGURE 4.11
Sample of Low-Level Revisions for Ben



Correspondingly, there were no increases in Ben"s lexis and grammar scores for writing tasks 1 and 2. When he did use Criterion feedback for writing task 3, both his scores for lexis and grammar increased. The scores for writing task 1 reflect Ben"s focus on developing ideas and organizing his ideas to be more logical. His scores for task response and organization increased. Likewise, the lack of change in scores for writing task 2 corroborates him being too busy with other commitments to develop his ideas and organization. Table 4.48 summarizes Ben's scores for task response, organization, lexis, and grammar for the first and the last drafts.

TABLE 4.48
Ben's Scores for First and Last Drafts for Task Response, Organization, Lexis, and Grammar

	Writing Task 1		**Writing Task 2**		**Writing Task 3**	
	First	Last	First	Last	First	Last
Task Response	2.5	3	3.5	3.5	3.5	3.5
Organization	3	3.5	3.5	3.5	3.5	3.5
Lexis	3	3	3	3	3	3.5
Grammar	3	3	3.5	3.5	3.5	4

In Ben"s new pieces of writing, after a significant increase of errors after pretest, the number of errors found by Criterion fluctuated for all three types of errors, with the delayed posttest having the greatest number of total errors (32). Table 4.49 reports the number of errors in new writings in mechanics, grammar, and usage from Criterion.

TABLE 4.49
Number of Errors in New Writings in Mechanics, Grammar, and Usage from Criterion for Ben

	Mechanics	**Grammar**	**Usage**	**Total**
Pretest	2	8	9	19
Writing Task 1	7	6	17	30
Writing Task 2	8	2	5	15
Writing Task 3	4	5	10	19
Delayed posttest	5	11	16	32

Ben reported that he had reservations against using automated feedback and that he mainly concentrated on developing his ideas and content. The increase in his scores seems to support this claim. His task response and organization scores increased between the pretest and delayed posttest (1.5 increase for both). However, his score for lexis and grammar only increased by 0.5. Table 4.50 reports the rater scores for new writings in task response, organization, lexis, and grammar.

TABLE 4.50
Rater Scores for New Writings in Task Response, Organization, Lexis, and Grammar for Ben

	Task Response	**Organization**	**Lexis**	**Grammar**
Pretest	2.5	2.5	2.5	2.5
Writing Task 1	2.5	3	3	3
Writing Task 2	3.5	3.5	3	3.5
Writing Task 3	3.5	3.5	3	3.5
Delayed posttest	4	4	3	3

The results of the Writing Process Questionnaire seem to indicate that Ben reported improving most for higher-level aspects of writing: he reported strongly agreeing to all statements in the writing phases except for the monitoring and revising at low-level. Unlike the other two students who strongly agreed with checking for grammar and vocabulary, Ben only agreed with the statements. This corroborates his assertion that he concentrated more on content and development than on grammar and lexis.

During the interview, Ben repeated that he does not believe that machines could give

effective feedback: "I don't think machines can do that. To correct your ideas." Like Rebecca, Ben was also surprised by the difference between Criterion scores and teacher feedback, which led him to be more distrustful of machine feedback: "I was shocked because I get almost a perfect score from the machine, and you tell me my paper is not good. Machine can't understand my writing."

Ben was also able to identify errors that Criterion did not find in his papers, which led to even more distrust and made him even more cautious of incorporating AWE feedback. In general, Ben did not find Criterion feedback useful for content and organization but found feedback on mechanics marginally useful. He was also ambivalent about continuing to use Criterion. During the delayed posttest, Ben reported that unlike the other two students, he did not use any automated tools while writing his assignments at university. Although Ben would have liked to seek out any writing help resources such as the ESL open learning centers, workshops, or the writing center for support, he did not do so because he did not have the time to attend these workshops and support services.

Overall, Ben"s overall view of machine feedback did not change during the treatment. In the interview and the Writing Process Questionnaire, he reported that he did not trust machine feedback and ignored it. In addition, when revising his work, he often missed his errors and introduced new ones. In the Writing Process Questionnaire, Ben reported that his writing process changed: he thought about his organization more and thought about how to make his ideas more persuasive. However, in the same questionnaire, he was only one of two students who reported checking his grammar less often than he did before the treatment. This may be due to his tendency to focus on development and organization and not trusting machine feedback. Likewise, in the Perception of Criterion Questionnaire, he reported that he did not find machine feedback useful for developing his writing. In addition, the analysis of his written work shows that when he revised, he focused on organization for better flow of his ideas but introduced more errors. Correspondingly, his scores for lexis did not improve during the treatment, and his grammar scores improved only marginally. In new pieces of writing, compared to the pretest, in the delayed

posttest, his writing became more developed and achieved a higher-level of cohesion and coherence. However, improvements in his lexis and grammar were only marginal.

Case 3: Jasmin

Class observations suggested that Jasmin was a very active student in the class. She routinely asked clarification questions, volunteered to answer teacher-initiated questions. This enthusiasm was also evident during teacher conferences. She would be taking notes on improvements, asking copious questions, and focusing mainly on surface-level errors because she wanted to make her essays "more perfect."

Jasmin reported that she had been in Canada for 17 weeks before the program started and has been accepted to the Finance Program at the university. Jasmin entered the program with the highest IELTS score compared to others in the experimental group. Her overall score was six and her writing score was also 6.

As seen from Table 4.40, Jasmin mostly agreed or strongly agreed that teacher feedback was helpful for revising her essays, and she reported that she pays the same attention to feedback and scores. She strongly agreed with the usefulness of computer feedback and used it often. She noted that she always revises her essays, and she enjoys revising them. However, like the other students, she reported that teacher feedback was not timely. She strongly agreed that revising is an essential part of writing.

For revisions, Jasmin submitted the greatest number of drafts in the experimental group. Seven for writing task 1, ten for writing task 2, and 9 for writing task 2. The changes in her revisions were all substantive, meaning that she made both high and low-level revisions. The effects of the revisions that she made were also all positive. Table 4.51 reports the number, magnitude of change, and effect of revisions between the first and the last drafts on writing quality for three writing tasks.

TABLE 4.51
Number, Magnitude of Change, and Effect of Revision between First and Last Drafts on Writing Quality for Jasmin

	Writing Task 1	**Writing Task 2**	**Writing Task 3**
Number of Revisions	7	10	9
Magnitude of Change in Revision	Substantive	Substantive	Substantive
Effect of Revision	Positive	Positive	Positive

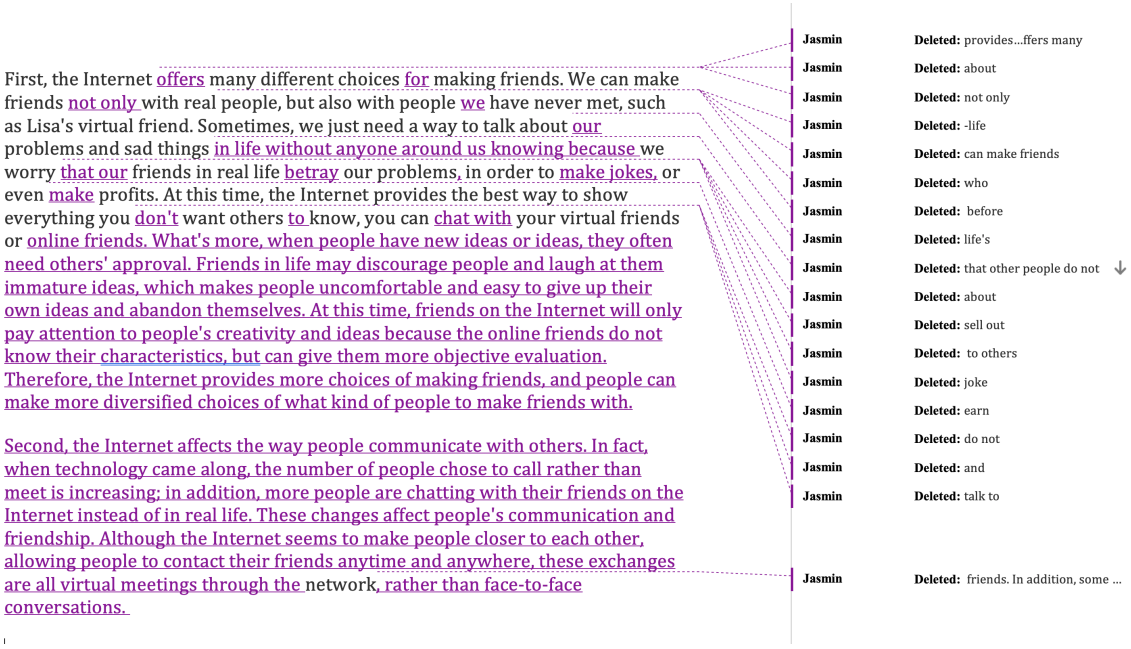
During her revisions, Jasmin strived to resolve all errors identified by Criterion; on average, she resolved 75 percent of all errors. She submitted a large number of drafts after addressing as many errors as possible. Table 4.52 summarizes the number of mechanical, grammar, usage, and total errors by writing tasks for the first and the last drafts.

TABLE 4.52
Number of Errors in First and Last Drafts in Mechanics, Grammar, and Usage from Criterion for Jasmin

	Writing Task 1		**Writing Task 2**		**Writing Task 3**	
Revision	**First**	**Last**	**First**	**Last**	**First**	**Last**
Mechanics	3	1	0	2	1	2
Grammar	3	0	2	0	4	3
Usage	13	1	8	2	10	0
Total	19	2	10	4	15	5

The analysis of change in revisions showed that Jasmine made significant changes between drafts and made revisions to resolve the corrective feedback generated by Criterion. In writing task 1, as seen in Figure 4.12, Jasmin resolved all errors identified Criterion, but added new detail in the first paragraph to elaborate her ideas and added a second paragraph to make her ideas flow better.

FIGURE 4.12
Sample of High- and Low-Level Revisions for Jasmin



Similarly, in writing task 3, she made several changes to help elaborate and develop her ideas more fully. For example, she wrote, "Firstly, many students choose single-gender schools because of their region." which she revised to "First, single-sex schools only have boys or girls, so there is no gender difference and stereotype, so students can be more confident and self-respecting in school life and study." The revised draft elaborated her idea and explained what she meant by region in the first draft.

Correspondingly, Jasmin"s scores for writing task 3 increased significantly between the first and the last draft, and her scores increased for all categories. Likewise, for writing task 1, her score for task response increased due to her revised paragraphs developing her ideas more fully. Like the other students, there were minimal revisions made for writing task 2. This may have been due to the midterm tests that took place in the same week. Table 4.53 summarizes Jasmin's scores for the first and the last drafts for task response, organization, lexis, and grammar for the first and the last drafts.

TABLE 4.53
Jasmin's Scores for First and Last Drafts for Task Response, Organization, Lexis, and Grammar

	Writing Task 1		**Writing Task 2**		**Writing Task 3**	
	First	**Last**	**First**	**Last**	**First**	**Last**
Task Response	2.5	3	3	3	3	4
Organization	3.5	3.5	3.5	3.5	3.5	4
Lexis	3	3.5	3.5	3.5	3.5	4
Grammar	3	3.5	3.5	3.5	3.5	4

In Jasmin"s new pieces of writing, after a significant increase of errors after pretest, the number of total errors found by Criterion decreased steadily from 19 for writing task 1 to 8 for the delayed posttest. Specifically, Jasmin decreased errors for grammar and usage and her error rates for mechanics stayed similar. for all three types of errors, with the delayed posttest having the greatest number of total errors (32). Table 4.54 reports the number of errors in new writings in mechanics, grammar, and usage from Criterion.

TABLE 4.54
Number of Errors in New Writings in Mechanics, Grammar, and Usage from Criterion for Jasmin

	Mechanics	**Grammar**	**Usage**	**Total**
Pretest	0	0	2	2
Writing Task 1	3	3	13	19
Writing Task 2	0	2	8	10
Writing Task 3	1	4	10	15
Delayed posttest	4	1	3	8

Jasmin reported that she was very engaged with both teacher and machine feedback. The increase in her scores seems to support this claim. Her task response and organization scores increased between the pretest and delayed posttest (1.5 increase for both). Similarly, her scores for lexis and grammar increased by 1 between the pre- and delayed posttests. Table 4.55 reports Jasmin"s scores for new writings on task response, organization, lexis, and grammar.

TABLE 4.55
Rater Scores for New Writings in Task Response, Organization, Lexis, and Grammar for Jasmin

	Task Response	**Organization**	**Lexis**	**Grammar**
Pretest	3	3	3	3
Writing Task 1	2.5	3.5	3	3
Writing Task 2	3	3.5	3.5	3.5
Writing Task 3	3	3.5	3.5	3.5
Delayed posttest	4.5	4.5	4	4

The results of the Writing Process Questionnaire for Jessica did not change much from the initial questionnaire before the treatment. This is because she responded "strongly agree" to almost all questions in both pre and post-treatment questionnaires.

During the interview, Jasmin suggested that she viewed teacher and Criterion feedback as complimentary. She saw both types of feedback as a resource to improve her writing. Although she found the teacher conferencing intimidating at first, she found the one-on-one interaction motivating. She also acknowledged that Criterion was not capable of always finding or giving accurate feedback but was still a helpful source of feedback when combined with teacher feedback. Like Ben, she does not trust Criterion: "I don't believe in technology," but she was much more engaged with Criterion feedback because Criterion "is useful because I can fix my some grammar mistakes." In addition, she saw resolving errors as a game: "I try to find all the mistakes so there are no highlights. It is like a game," which encouraged her to engage with the feedback more. Even though she did not trust AWE feedback completely, she still saw value in the feedback. Also, unlike the other two students, Jasmin found the generic feedback and highlights helpful in thinking about content and development more deeply: "sometimes I think why my idea is good for this topic, and it will give me the feedback and let me know what and where it is, so and it also changed my mind about the topic." In general, Jasmin reported that she found the various types of feedback that Criterion offered to be helpful, and she would continue to use it in the future.

During the delayed posttest, Jasmin mentioned that she did not have the time to revise her essays due to the sheer number of writing assignments she had in her university courses after the EAP program. Unlike the other two students who only had one or two essays like assignments in their courses, Jasmine had more than seven papers. She reported continuing to use her word processor's grammar and spell check function but did not seek out any writing help resources in the library such as the ESL open learning centers, workshops, or the writing center for support due to lack of time.

Overall, Jasmin found the integration of Criterion and teacher feedback very helpful for revising her essays, and she had a positive view of the integration of both types of feedback. In the interview and the Writing Process Questionnaire, she reported that although she found the generic feedback frustrating, she overall found the feedback useful for noticing her errors which gave her a chance to revise and resolve them. She also found the game-like element of hunting down the errors satisfying and consistently made substantive changes. However, in the Writing Process Questionnaire, Jasmin reported no difference between the pre-and the post-treatment. This may have been because she selected strongly agree for most questions before and after treatment. Jasmin also responded very favourably to the Perception of Criterion Questionnaire except for its scoring aspects. The analysis of her written work shows that she concentrated on content and development. In revisions, she added full paragraphs for development and added signposts to help the reader follow her ideas. In addition, she often resolved most errors in mechanics and grammar between revisions, and her errors steadily decreased in new writings. In new pieces of writing, compared to the pretest, in the delayed posttest, she received the highest rater scores in the group for task response and organization, reflecting her constant effort in connecting ideas and developing her ideas. Moreover, her lexis was more sophisticated, and she made fewer grammatical, mechanical, and usage errors.

The following chapter summarizes the key findings of the study and discusses them.

Chapter 5: Discussion and Conclusion

This chapter summarizes and discusses findings concerning the effects of the use of hybrid corrective feedback and evaluation on students' L2 academic writing practices in a post-secondary ESL writing classroom. At the end of the chapter, some initial pedagogical and theoretical implications are discussed, and future research areas are identified.

5.1 Summary and Discussion of Findings

The purpose of this quasi-experimental study was to measure the effects of combining feedback from an automated writing evaluation (AWE) program, Criterion, with teacher feedback in an intensive EAP writing class preparing students to enter an undergraduate degree program at a large university in Canada. Firstly, to determine if there were any differences in changes in approach to writing between the comparison and experimental groups, students were asked about their writing processes before and after the treatment using a Questionnaire. The findings show that, firstly, both groups of students changed their orientation toward writing to attend more to some writing processes, possibly due to the learning effects. After the treatment, the students in the experimental group reported thinking more about the correct usage of grammar to express their ideas, how to connect ideas more smoothly, and the logical order of ideas in the generating texts phase than did the comparison group. In addition, the experimental group students reported monitoring and revising the coherence of their texts, checking for the possible effect of their writing on the audience, and checking their grammar and lexis more often than did the comparison group after the treatment. Furthermore, students tended to make surface-level changes and made significant changes to the content, organization, and development of their essays after the intervention than they did before, as shown by the ratings of the magnitude and change of revisions.

Secondly, to examine changes in the language, content, and organization of students' writing due to hybrid feedback, pretests and delayed posttests were conducted. Comparisons were also made between the experimental and comparison groups and between the first and the last drafts for in-class writing tasks for the experimental group and in new pieces of writing. Both groups improved the content, organization, and language of their writing. The drafts improved in terms of writing quality, and there were marginal differences between the comparison and the experimental groups for new pieces of writing. Lastly, to examine students' perceptions of Criterion and teacher feedback, questionnaire and interview data were gathered to obtain more insights into students' engagement with and perceptions of hybrid feedback to explain and elaborate on some possible causal links between perception, engagement, and utilization of feedback. The questionnaire and the focus-group interview findings indicated that the students in the experimental group generally held positive attitudes toward using AWE to improve their writing while signalling reservations and critiques about the fallibility of AWE feedback, the usefulness of generic feedback, and frustrations with scoring. On the whole, students found Criterion feedback helpful but found the combination of Criterion and teacher feedback to be more specific and contingent. The following sub-sections discuss the findings of the study related to the research questions in more detail.

5.1.1 The Effects of Hybrid Corrective Feedback on Students' Approaches to Writing

In general, both groups' approaches to writing became more process-oriented, but students in the experimental group reported becoming more process-oriented than the comparison group: the students who received hybrid feedback seem to see writing more as a process. This finding aligns with previous research on feedback that provides opportunities for revising. Lee (2017) found that feedback that provides AaL in the writing classroom facilitates process-orientation to the writing. However, the results of this study showed that integrating teacher and automated feedback in a hybrid feedback system seems to have changed students' perception of writing from a product to a process

as reported by the students. This is supported by the number of revisions submitted by the experimental group and by the interview and questionnaire data revealing that the students in this group were motivated to write more. In a recent study that examined student writing motivation in Hong Kong, Lee et al. (2018) found that students' low motivation to write may be due to a focus on the written product; that product-oriented feedback tended to be demotivating. Process-oriented feedback that gives students multiple opportunities to revise and draft, as was the case in the current study, seems to increase students' motivation to write (Duijnhouwer et al., 2012).

For changes in writing processes, the questionnaire data indicated that the experimental group specifically improved in the following three phases: generating texts, monitoring and revising at a high level, and monitoring and revising at a low level. This finding is consistent with previous research, which reported that AWCF facilitated more revisions and higher motivation to write (El Ebyary & Windeatt, 2010; Li et al., 2014; Warschauer, 2010; Zamin, 2021). The increase in the number and quality of revisions in this study suggests that integrating AWE in a hybrid manner encourages students to revise more, reflecting similar findings in previous research (Grimes & Warschauer, 2010; Warschauer & Grimes, 2008; Warschauer & Ware, 2006) but contradicting some research that found a low uptake of AWE feedback and a lack of motivation to revise among students receiving AWE feedback (Attali, 2004; Bai & Hu, 2017; Li et al., 2015; Tian & Zhou, 2020). The students in this study made an average of 4.25 revisions for each writing task. Although previous research has noted that there may be a tendency for students to use AWE tools less as time passes due to lack of motivation (Zhu et al., 2020), in this study, the opposite seems to have happened. While it is true that the number of revisions for writing task 3 was 3.71, this may have been due to two reasons. First, the last two weeks of the class are "crunch" time, when all assignments, including the major writing assignment and the research paper, are due, final tests for reading and listening are administered, and final presentations are evaluated. Therefore, the marginal decline in the number of revisions may indicate that the students may have been more motivated to revise even when they are very busy. Second, for writing task 1, there were seven substantive changes, 9 for

writing task 2, and 9 for writing task 3. This shows that the magnitude of revisions did not decline across tasks. This pattern contradicts the finding in the literature that students' motivation to revise falls as the novelty effect of using AWE feedback declines (Sung et al., 2016). In addition, the selected individual cases show that students who had positive attitudes towards automated feedback made more substantive revisions, wrote more drafts, and continued to use similar tools after the treatment. This is supported by Zhang and Hyland (2018), who found that highly engaged learners became more autonomous in their learning because the immediacy of AWE feedback was conducive to deeper engagement. The engagement with the feedback in this study may have been due to its hybrid nature. When engaging with only AWE feedback, students may not be able to resolve the errors due to not understanding the metalinguistic feedback or because they believe that machines do not understand their writing (Chen & Cheng, 2008); however, with the hybrid approach, the students know that their writing will be read by a human and the teacher can mediate machine feedback (Wang, 2013).

While previous research focused on the unfocused, generic, non-dialogic, and fallible nature of AWE feedback (see Jiang & Yu, 2020), with this study's implementation of hybrid feedback, the reported increase in the experimental group's monitoring and revising processes suggests that the combination of AWE and teacher feedback may have helped the students develop the skills to become "autonomous students" who can "draw on various sources of knowledge [and resources] to strategically respond to AWE feedback" (Bai & Hu, 2017, p. 79). This is in contrast to previous studies that examined the effects of AWE feedback only (Chen & Cheng, 2008; Grimes & Warschauer, 2010; Li et al., 2015). The increase of self-regulatory monitoring and revision strategies for the experimental group compared to the comparison group after the treatment suggests that combining teacher and automated feedback can encourage students to notice and reflect on their writing, which helps students refine and improve their self-editing skills; a finding that is consistent with current WCF literature (see Bitchener & Storch, 2016). Furthermore, the results from the Writing Process Questionnaire confute previous criticism of the use of AWE in the classroom for corrective feedback as being overly

prescriptive and relying only on surface-level errors. The opposite was found for hybrid feedback in this study as evidenced by the ratings of the revisions the participants in the experimental group made and their responses to The Writing Process Questionnaire as well as results from the individual case analyses (c.f. Grimes, 2008). The significant increase in monitoring and revising at a high-level for students in the experimental group indicates that students paid more attention to the content and organization of their essays, which disputes assumptions that the use of AWCF overemphasizes and will lead to a focus on surface-level errors. The questionnaire data indicated that after receiving feedback from both the teacher and AWE, the experimental group increased their monitoring and revising, suggesting that AWE feedback may promote learner autonomy (Stevenson, 2016).

Overall, students in the experimental group changed their orientation to writing more than did the comparison group. Not only that, but the experimental group reported focusing more on high-level revisions such as having a strong viewpoint, content development, and organization. In addition, the change in the experimental group's orientation to writing from product to process-oriented and in their level of engagement, as evidenced by the number of revisions they made, suggest that the process-oriented nature of hybrid feedback may have had a more significant impact on students' motivation and engagement with the feedback. For instance, Harks et al. (2014) noted that process-oriented feedback had a greater positive effect than grade-oriented feedback on changes in students' achievement and interest. Moreover, the greater increase among the different cognitive phases of the writing process for the experimental group compared to that for the comparison group suggests that the combination of feedback helped reorient students' views of writing from product to process. The experimental group reported thinking more about correct sentence structures, logically connecting ideas, and content development. The interview data supported these findings: there was a consensus among the interviewees that the combination of feedback helped them employ metacognitive strategies to monitor and revise texts on which they had received automated feedback. These findings correspond with language learning strategy theory

(Oxford, 1990), which states that metacognitive aids help focus learner attention on recurring errors and empirical studies that show a positive correlation between metacognition and writing development (Bitchener & Storch, 2016).

5.1.2 Changes in Written Products

In new pieces of writing, the analyses of fine-grained indices showed a general improvement over time for both groups indicating a learning effect but did not show a significant interaction effect for group with time. In other words, while both groups made significant gains, there were no significant gains due to treatment. For example, for both groups, all indices improved between week 1 and 3 months after the program except for lexical depth, local cohesion, and global cohesion.

Although the analyses of fine-grained indices did not reveal any significant interaction effects for group with time, analyses of human ratings detected significant effects on the quality organization, lexis, and grammar. There were significant effects on scores for new pieces of writing for group and time, indicating that their writing scores improved in subsequent writing tasks. However, there were no significant interaction effects for group with time, except for grammar ratings, which were associated with a weak significant interaction effect. The grammar scores in the experimental group improved more than the comparison group. This may have been due to a combination of reasons. First, the different environments in which the delayed posttest was completed due to the pandemic may have influenced the quality of the essays. Second, research has indicated that students with lower proficiency improve their lexical and grammatical range and accuracy before they improve topic and development (Beers & Nagy, 2009), which may explain why the experimental group, which had lower proficiency, made more gains in terms of grammar ratings but not on ratings of other writing features. Third, human raters may account for topic development and content when rating organization and lexis, which the automated indices do not consider. For example, while the index of word length is seen as an indicator of lexical frequency – longer words being less frequent, previous studies have shown no correlations between human judgements of lexical

frequency with indices of word length (see Crossley et al., 2011) but human judgements of lexical frequency were highly correlated with topic development (Bitchener & Storch, 2016).

In contrast, when comparing the first and the last drafts on the same task for the experimental group, the analyses indicated that the last drafts exhibited gains in most of the indices examined. Specifically, every measure showed improvement on the last draft compared to the first draft, except for global syntactic complexity and lexical frequency. This finding may indicate that, while students corrected syntactical errors, they did not write more complex forms and that, while the range and depth of their lexis increased, lexical frequency did not change. Similarly, for human ratings of writing quality, while there were significant effects for drafts and time, there were no significant interaction effects for drafts with time, except for grammar. The grammar scores in the experimental group improved more than the comparison group. The following paragraphs summarize the main changes in indices for fluency, lexical complexity, organization, and quality of response for new pieces of writing and differences between the first and the last drafts.

For changes in fluency, there was no increase in the number of words in new pieces of writing, but there was a marked increase between the first and the last drafts. Contrary to previous findings that AWE feedback led to more extended essays (Grimes, 2008; Schroeder et al., 2008; Warschauer & Grimes, 2008), the students in the experimental group in this study did not write more extended essays compared to the comparison group when responding to new tasks. This could be explained by the timed nature of the first revision due to the constraints of the curriculum. The length of all essays was around 500 words, which is the required length for in-class writing tasks. Students may have produced the required number of words and focused more on other areas of writing, such as organization, grammatical accuracy, and lexical sophistication. However, when comparing drafts on the same task, students wrote more in the last drafts than they did in the first draft. This seems to corroborate the finding, from the case analyses, interview, and questionnaires data, that students focused on content and development when revising.

The analysis of students' last drafts indicated that they elaborated more by giving more details and developed their ideas more fully. The findings support previous research on the revision behaviours of students utilizing automated feedback, which indicates that these students tend to write longer when they revise their texts (Kellogg et al., 2010; Warschauer & Grimes, 2008).

For syntactical complexity, the most significant gains in new pieces of writing for both groups were made in the degree of phrasal sophistication and the number of NP types in new pieces of writing. In new drafts, students in the experimental group made significant gains in NP types as expected from findings from the previous studies that examined syntactic sophistication and complexity (Dikli, 2006; Hoon, 2006; Ware & Warschauer, 2006). Specifically, in the last drafts, students included more prepositional phrases in their essays; prepositional phrases are a key indicator of academic writing (Biber & Gray, 2010). A recent study by Casal and Lee (2019) found that human raters scored essays that included significantly more prepositional phrases higher than essays with fewer prepositional phrases.

For syntactical accuracy, while there were no significant differences between the experimental and the comparison group for new pieces of writing and between the first and the last drafts, there were accuracy gains for both. Although the experimental group students stated that they focused more on grammatical accuracy than the comparison group, the findings show that the gains were less than expected. However, in drafts, it was clear that students used automated feedback for revising and resolving errors. Three possible interpretations may explain the limited gains in accuracy. The first is that it may be easier to acquire complexity than accuracy because students may not have reached the level of automatization of specific grammatical rules and are less likely to benefit from focusing on grammar rules beyond their current level of development (Fukuta et al., 2019). The second possibility is the U-shaped course of development of SLA, which proposes that some students may have limited success in acquiring accurate forms because, after initial exposure to corrective feedback, they may use the correct forms but

may regress in subsequent writing before internalizing the correct forms (Ellis, 1997). This could occur because integrating and internalizing knowledge and skills are gradual and incremental (DeKeyser, 2007), and multiple opportunities for noticing gaps and practices are required for deep processing and internalizing noticed language forms (Schmidt, 2001). Students with lower proficiency may take a longer time between when the procedural knowledge is acquired and when they can demonstrate its use – this is especially true with more advanced grammatical forms (Bitchener & Storch, 2016). This may explain why the experimental group, which was less proficient in more advanced grammatical forms, did not achieve significant gains in syntactic accuracy. Studies on the effects of AWE feedback on grammatical accuracy seem to support this hypothesis (Feng et al., 2015; Li et al., 2017). The third possible reason may be the limited recall of Criterion for detecting errors. Past research shows that automated detection of errors may have higher precision but may have lower recall. Precision is the ratio of the number of relevant writing features retrieved to the total number of irrelevant and relevant writing features retrieved (in percentage), while recall is the ratio of the number of relevant writing features retrieved by the engine to the total number of relevant features retrieved by a human marker (in percentage) (Link, 2015). In other words, errors that are identified by the AWE are generally accurate, but the automated systems failed to detect many errors (Crossley et al., 2019a; Hoang & Kunnan, 2016). Therefore, the analysis based on Criterion data for the number of errors may not have been complete. In contrast, the human ratings for grammar saw significant increases in scores for group and time. They also saw a weak significant interaction effect for group with time, suggesting that human raters may be more sensitive to grammatical accuracy due to greater recall and precision than Criterion.

For lexical complexity, lexical frequency scores fluctuated. By the posttest (week 7), the experimental group decreased frequency scores, indicating more usage of complex lexis, but in the delayed posttest (3 months after treatment), the experimental group's score increased, indicating usage of less complex lexis in new pieces of writing. Lexical range increased for both groups for function words and bigrams, indicating that the given text

contained a more comprehensive range of words demonstrating greater lexical sophistication. However, there was no significant difference in lexical depth, and there was only a marginal increase for lexical frequency scores and range scores remained similar between the first and the last drafts. Crossley et al. (2014) reported that the strongest predictor of essay quality was the number of word types used by an L2 writer. The increased range of lexis for the experimental group in new pieces of writing gives more credence to their findings, which by and large aligns with previous results (Lu, 2011; Wolfe-Quintero et al., 1998; Wolfe-Quintero et al., 1998). There may be several reasons for the findings. Feedback for lexical frequency, range, and depth is not rule-based, which means that AWE cannot give specific feedback for improvements due to the limitations of the current implementation of AWE metalinguistic explanation. AWE feedback only gives generic feedback for word choice and collocation, which may have resulted in less uptake from the students. Also, the lack of improvements for lexical depth may be because it is conceptually easier to increase the range. For example, students can use the thesaurus function in their word processors or use online dictionaries to increase range. However, increasing depth is more difficult to acquire because the students need to understand the specificity of the words. As reported in the interview, students found the flagging of repeated words and phrases by Criterion challenging to resolve. However, the data suggest that flagging common words, collocations, and easily confused words by Criterion facilitated students to increase the range of academic vocabulary. The increase in frequency score may be due to the explicit teaching of academic vocabulary in the class, which may have resulted in higher usage in students' writing resulting in higher frequency scores.

For coherence, in new pieces of writing, there was a lack of significant interaction effects of time by group on fine-grained cohesion indices, and in drafts, students repeated key nouns more often, and there were higher incidences of conceptual overlap of nouns between paragraphs in the last drafts than the first drafts. Examinations of micro features for cohesion in previous studies have resulted in mixed findings. In some studies, the presence of explicit cohesive devices was associated with higher quality writing (Jin,

2001), but in other studies, the presence of cohesive devices such as lexical overlap, semantic overlap, givenness, and connectives was associated with lower quality (Crossley & McNamara, 2012; Guo et al., 2013). In the current study, both groups made minor gains in local and global cohesion, but both groups made major gains in text cohesion, which is one of the main focuses of the curriculum. The program's explicit instruction in using connectives, transition sentences between paragraphs, and repetition of keywords most likely affected the observed increase in text cohesion for both groups. Similarly, the increase in repeated content words and incidence of conceptual overlap of nouns between paragraphs between the first and the last drafts for the experimental group likely reflects the curriculum. In previous research, repetition of key nouns or noun referents was shown to be a key indicator of text cohesion (Crossley et al., 2016; Kyle et al., 2016; McNamara et al., 2014).

Human ratings of task response, which measures content and development, increased across new writing pieces for both groups and between the first and the last drafts for the experimental group. The statistical analysis detected significant differences between the two groups. While the experimental group's pretest showed lower scores than the comparison group for task response, by the delayed posttest, the experimental group's scores increased more than those of the comparison group, perhaps reflecting the changes the students reported in their writing processes. Students in the experimental group reported thinking more about development, content and how their essays would be perceived by the reader than did the comparison group. These results are consistent with the experimental group's perception that they tended to revise for relevant content in their writings. The results concerning task response for both new pieces of writing and drafts are contrary to previous research. For example, Warschauer and Grimes (2008) found that students made no significant changes in their drafts and mainly focused on spelling, punctuation, and grammatical errors. However, the findings of this study suggest that the combination of AWE and teacher feedback helped students in the experimental group attend to organization and content as much as, or even more than, the mechanical aspects of writing since students knew that ultimately, they would have a human audience for

their writing. In addition, the hybrid feedback seems to have mitigated one of the more significant failings of AWE feedback: failure to address content and development (Attali & Burstein, 2006; Dikli & Bleyle, 2014).

Overall, the findings of this study show that students in both groups increased their writing quality. In general, both groups' writing quality improved due to the learning effect in new pieces for writing: scores for task response, organization, lexis, and grammar increased for both groups, but the scores of students in the experimental group increased more in new pieces of writing. For changes between the first and the last drafts for the experimental group, in general, students made improvements in all indices across drafts, with scores for task response increasing both in new pieces of writing and between the first and the last drafts, which suggests that the combination of AWE and teacher feedback helped students to focus more on content and development.

This study has added to a better understanding of the effects of combining teacher and AWE feedback on the development of academic writing in an intensive academic English ELL writing program and the value of hybrid feedback for students, as well as the limitations and constraints of machine mediated feedback and scoring. The inclusion of the Writing Process Questionnaire, the interview, and the case analyses, in addition to fine-grained indices and scores, provided greater clarity by providing data about students' perceptions of changes in their writing processes. Automated indices may not have detected the underlying changes in students' writing processes because changes in development, content and some aspects of organization may have been undetectable due to Criterion and rater scores not being sensitive to changes in writing development.

Both the qualitative and quantitative findings have contradicted findings in earlier research that the use of AWCF in the classroom results in mechanical and superficial changes in students' revision and writing behaviour (see Li et al., 2015). For example, data from the rating scale for revision indicated that students in the current study tended to make as many changes for content and organization as for grammar and lexis between

the first and the last drafts. Accordingly, the Writing Process Questionnaire data indicated that the treatment had a statistically significant, but weak, effect on students' overall monitoring and revising behaviour at low-level (uptake) and had a significant strong effect for improvement in new pieces of writings (retention). An examination of the automated fine-grained indices showed marked changes for syntactic sophistication, complexity, and accuracy in new writing pieces for students in the experimental group compared to students in the comparison group, indicating retention of more sophisticated and complex forms and increased accuracy. In further analysis, the analysis of individual questions in the Writing Process Questionnaire showed that students did focus on syntactic sophistication and complexity along with accuracy. The greater gains in sophistication and complexity in new pieces of writing, compared to revisions, may indicate that students seem to have noticed and internalized the syntactic structures due to the automated feedback they received during the revision process. Moreover, the increase in accuracy for both new pieces of writing and across drafts on the same task may indicate that specific feedback on grammatical accuracy is easier to resolve than generic feedback for word choice and repetition of words by Criterion. This finding is consistent which the finding in some previous studies that student success in revising translates into an improvement in new writing, but there may be variation across individual students and error types (Chandler, 2003; Ferris et al., 2010).

5.1.3 Students Views of Hybrid Corrective Feedback

In general, the students held a positive attitude toward using hybrid feedback: they found the instantaneous feedback of Criterion useful and were motivated by an additional source of feedback that complemented teacher feedback. However, the students paid more attention to and preferred teacher feedback than AWE feedback due to it being more specific and dialogic to their needs, and because it helped with clarifying feedback, they could not resolve from Criterion. The combination of feedback helped students feel more ownership of the learning process because while teacher feedback is seen as authoritative, the students felt they had agency over AWE feedback. Overall, the experimental group agreed with the usefulness of the feedback, especially on instantaneous feedback and

feedback on grammar, usage, and mechanics. However, they felt more neutral about feedback on style and organization & development. Student overall engagement and satisfaction with automated feedback may have been because of its immediacy (Polio, 2012). Zhang and Hyland (2018) found that for highly engaged learners, AWE feedback promoted a more autonomous engagement with feedback than did teacher feedback due to the immediacy of AWE feedback, which supports previous findings that immediate feedback is more conducive to deeper engagement (Bitchener & Storch, 2016). The questionnaire and interview responses suggest that students were motivated by the additional feedback from Criterion to complement teacher feedback, a finding that is consistent with those of other classroom-based studies (Chen & Cheng, 2008; Dikli & Bleyle, 2014). However, during the interview, some students questioned the accuracy and, therefore, the usefulness of Criterion feedback, reflecting their awareness of Criterion feedback limitations. While Stevenson and Phakiti (2014) noted that students' engagement with AWE feedback does not necessarily mean that they have adopted cognitive and metacognitive strategies to notice, evaluate, and finally improve their writing, the Writing Process Questionnaire data show that engagement with AWC feedback combined with teacher feedback appears to have changed students' writing processes to some extent.

Another point to note is that while students found AWE feedback helpful, only just over half expressed satisfaction with Criterion. Their dissatisfaction may be because of the difficulties they experienced interpreting the scores that Criterion assigned to their essays and the user experience of Criterion. As previous research has indicated, students found the holistic and trait scores of Criterion problematic (Link et al., 2014). The results of this study indicated that students were concerned with how the scores were calculated and reported that the Criterion scores were inadequate in providing an understanding of their performance. Students' perception that automated scores were not useful somewhat invalidates the claims of some previous studies that these scores could be helpful in the classroom as a summative assessment tool. There may have been two reasons for this. Firstly, previous research suggests that because the scoring engines of AWE systems do

not measure the construct in the same way as humans do, but on feature weights based on predicting human ratings, the scoring of AWE may not necessarily reflect the relative importance of what human raters deem most important in writing (Attali, 2015). Secondly, the lack of variability of Criterion trait scores and lack of transparency for its holistic scores may have confused the students more about their writing performance.

Students reported preferring explicit feedback to generic feedback from Criterion because it was easier to understand and resolve. This may have been because Criterion metalinguistic explanations do not change or adapt according to students' levels of L2 development. In other words, metalinguistic explanations treat errors of the same category by giving the same feedback regardless of the actual text. Further, students may not understand the abstract diagnostic language of generic feedback since it is divorced from context and does not provide examples or alternative models. Moreover, Criterion feedback does not provide a follow-up response to errors that the students failed to resolve in subsequent revisions. Criterion feedback "lacks cohesion and coherence for two reasons: first, it is not staged and purposeful, and second, the rationale is not regulated and delivered with regard to what feedback has already been provided to the student in previous submissions" (Mehrabi-Yazdi, 2018, p. 904). Some students gave examples of demotivation and frustration due to Criterion feedback, and quantitative data showed that some students paid very little attention to AWC feedback.

The interview showed that after using Criterion feedback, students seemed to appreciate and understand that the teacher's feedback was dialogic and tailored to each student's own needs. Although students stated that while they would prefer teacher feedback, they understood the time-consuming nature of WCF. The findings in this research support previous observations that students paid more attention to teacher feedback. However, in this study, the findings show that students exercised agency by adopting only the feedback they deemed useful and deciding if and how it would be incorporated in their revisions. This finding is consistent with previous research and is in contrast to studies that raised concerns that students would blindly resolve all automated feedback (Chen &

Cheng, 2008). Another explanation for why students ignored automated feedback may be due to students not trusting machine feedback because they felt the feedback they received was not accurate. In the individual case analyses, Ben, one of the participants, carefully considered and ignored AWECF because he believed that the feedback was incorrect, which stemmed from his dislike of writing to machines as "machines cannot correct your ideas." Therefore, while previous studies assumed that the non-use of the feedback was the result of students ignoring feedback (Lavolette et al., 2015; Ranalli et al., 2017), the current study gives credence to an alternative hypothesis that the non-use of feedback may be due to students noticing the feedback and ignoring it because they believe it is not accurate, that it is not necessary to resolve, and/or because they did not trust the source of the feedback, in this case, the machine. In other words, this study indicates that the hybrid feedback seems to have facilitated students' engagement of one of the central cognitive interactionist constructs, noticing, with different levels of awareness (Schmidt, 2001) to improve the quality of their writing.

The results above agree with those of other studies that AWE feedback promotes autonomy (El Ebyary & Windeatt, 2010), and positive outcomes ensue when students' autonomy is supported (Jang et al., 2010). The results are consistent with the findings from previous studies (Chen & Cheng, 2008; Dikli & Bleyle, 2014; Dikli, 2006; El Ebyary & Windeatt, 2010; Griffiths & Nicolls, 2010; Hoon, 2006; Shermis et al., 2008a; Warschauer, 2010; Yannakoudakis et al., 2018), which suggest that students are motivated by the game-like element of 'hunting' down errors to produce better texts due to the immediacy and accessibility of AWE feedback without the restriction of time and frequency of use.

While not a focus of the current study, a factor for students not being satisfied with Criterion may have been its user experience (UX). Students expressed that the UX of Criterion was "too busy," "not pretty," and the onboarding experience was not smooth. For example, a plethora of highlights and feedback that the students were initially presented with may have affected their perceptions regarding the complexity and overall

effectiveness of UX of Criterion: the students found the user interface overly busy and challenging to navigate. Research on UX design in enhancing student motivations in language learning, Seppälä et al. (2020) found that UX has a significant impact on creating deeper engagement and motivation when students engage with software for learning languages.

Overall, while some of the findings reflect those of past research, some results contradict previous research. As in previous research, the results of this study show that students felt more ownership of the writing process through AWE mediation and found feedback on mechanical aspects helpful. The findings also reiterate previous studies' assertions that students' perceptions and attitudes towards WCF have significant effects on their writing and revision practices (Bitchener & Storch, 2016; Lee, 2005; Lee, 2008a). This may have resulted in some students giving little value to AWCF. The differences in students' perceptions and preferences of WCF were revealed in the questionnaire. The interview data suggest that students who have a positive perception of WCF and, by extension, AWCF were more positively affected by CF, AWCF, and hybrid feedback reflecting findings from previous research that students' engagement with AWE feedback and the effects of AWE feedback on writer's revision strategies and writing processes are mediated by students' perception of the tool (Bai & Hu, 2017; Cheng, 2017; Link et al., 2014; Warschauer & Ware, 2006). Also, as expected, based on previous research, students preferred explicit rather than generic feedback because machine feedback is divorced from context and students' ZPD. Aljaafreh and Lantolf (1994) suggested that effective intervention should be graduated, contingent, and dialogic. Research has shown that graduated WCF can be more effective than non-graduated WCF (Nassaji & Swain, 2000; Rassaei, 2014). In other words, too little assistance is undesirable, whereas too much assistance may be harmful (Lee, 2017). In contrast to studies that raised concerns that students would blindly accept automated feedback, the students in this study tended to evaluate feedback and disregard feedback that they felt was unhelpful or incorrect while they saw teacher feedback as authoritative. When students had doubts about machine feedback, they asked for clarification in class or during the oral conference. This suggests

that teacher mediation may be necessary and encouraged to help students understand AWE feedback's limitations.

5.2 Limitations

Despite the best efforts to conduct the current study, some limitations may have an impact on the validity of its conclusions. This section details these limitations then makes recommendations that can be addressed in future research.

One major limitation of the study is that different tasks were used on different occasions, thus confounding time and task effects on the participants' writing performance. For example, the varying difficulties between the tasks and the different kinds of vocabulary the tasks elicited may have affected student performance. Although the way hybrid feedback was implemented in the current study showed positive effects on students' writing processes and the quality of their academic writing in post-secondary ELL contexts, it may not be the best way of implementing hybrid feedback. There may be other combinations of AWE and teacher feedback that may yield better results. Some researchers suggest that students submit their writing to the AWE system first until a threshold score is reached, then submit the last draft to the teacher (see Chen & Cheng, 2008). This was the approach adopted in this study. However, other researchers have suggested that restricting access to AWE may change the outcome of the study because they found that students engaged with AWE feedback more when they had limited access to it compared to when they had unlimited access to machine feedback (Hoang & Kunnan, 2016). Also, the length of treatment, the approach to mixing AWE and teacher feedback, and the focus of each type of feedback may be varied to gather empirical evidence to determine the best way to combine AWE and teacher feedback for different classroom contexts.

This study's most significant limitation was the low sample size in the experimental and comparison groups (N=16 and 17 respectively). The low sample size may have limited

the findings because the smaller the sample, the more difficult it is to detect significant effects (Stevenson & Phakiti, 2019). Another limitation is the duration of the study. Although the students were enrolled in an intensive 8-week program with 4 hours of instruction per day over five days a week, an eight-week period may be insufficient to observe significant changes in writing performance (Mazgutova & Kormos, 2015; Storch & Tapper, 2009). Although designed to be longitudinal, the current study only captured three snapshots of students' performance over eight weeks; the limited number of feedback sessions employing hybrid feedback may not be enough to observe its potential effects. Third, the specific population was international mainland Chinese students in an intensive EAP program with conditional acceptance. As a result, the findings may be difficult to generalize to other ESL programs with different demographics and different pedagogical orientations to language learning or different institutional purposes. For example, Bridgeman et al. (2012) indicated that students from mainland China tend to get substantially higher scores from Criterion than other student populations, perhaps due to their mother language or the education system that emphasizes formulaic bundles in writing (Qin, 2014). Additionally, the gatekeeping nature of the program where the study took place may have resulted in higher engagement with hybrid feedback.

The study examined the effect of the particular implementation of hybrid feedback on writing development by examining changes through the Writing Process Questionnaires, Criterion Perception Questionnaires, and focus group interview. However, the reliance on self-reported data may be problematic because of response biases (2014). Response bias may have impacted the results for research question 3 about students" views of hybrid AWE corrective feedback. Because the other research questions were augmented with other information to provide a more accurate picture of the results. Some researchers advocate think-aloud protocols to triangulate data, but the timed nature of in-class writing in the curriculum and the increase in cognitive load associated with think aloud protocols made it undesirable (Zamel, 1985). A concurrent protocol such as eye-tracking and videotaping of the writing session would have resulted in more robust multimodal data without causing significant interference during writing (Lim & Phua, 2019), but it was

not implemented due to a lack of resources.

In applied linguistics, the use of intact classes to provide a naturalistic setting for research is frequent, but this approach does not allow the randomization of group assignments, which may result in more confounding variables. Random assignments may reduce the influence of confounding variables because it distributes them at random between the groups (Adams & Lawrence, 2018). For example, the students in the experimental and comparison groups were not of equal proficiency. Other examples are student affective and external factors, such as anxiety, familiarity with technology, and family support. A random assignment of students to groups would have corrected for this and improved the study"s internal validity.

There may be two reasons for caution when interpreting the results of the delayed posttest. Firstly, the settings in which the delayed posttest was conducted for the experiment group differed from those of the comparison group due to the pandemic. Therefore, the results of the posttest can be questioned to some extent. However, I took care to monitor students in individual virtual rooms during the post test for the use of external tools and outside help. Nevertheless, the difference in the testing method may have affected students' motivation to write and had an impact on the results. Second, although this study's results revealed statistically significant improvement between the first and the last drafts, the improvement in new writing was marginal for both groups. One confounding variable may be the amount of writing and type of writing the students engaged in during the three months between the last in-class essay and the delayed posttest. In other words, the students' chosen majors in the university may have impacted their performance on the delayed posttest because the delayed posttest occurred about three months after their degrees started, and their writing practices may have been affected by the type of genre, the frequency of writing, and the writing tasks they had in their major classes.

Another factor that may have impacted the reported results is the lack of formal

evaluation of how proficient students were in using Criterion. However, the students received an onboarding session to establish that all students in the experimental group had a baseline proficiency with using the AWE system. I made efforts to answer questions and help students with Criterion use. In the final interview, while all students agreed that they engaged and were familiar with the tool, there may have been a differing level of proficiency using the tool, which may have impacted the students" engagement with the AWE system differently. A formal evaluation of proficiency with Criterion would have addressed this issue.

Despite these limitations, there were several strengths. The study is one of a few studies that have examined the hybrid approach to feedback rather than examining AWE feedback as a replacement for teacher feedback. Also, the use of an intact class and assignments that are part of the curriculum boosts the ecological validity of the study, which can enhance the generalizability and relevance of its findings to real-life classrooms compared to previous studies that only examined a very limited set of grammatical forms in laboratory settings. Although reliance on self-reporting may have been problematic, the analysis of macro and micro levels of changes in writing helped mitigate some of the biases of self-reported data. In addition, the in-depth analysis of the three selected cases helped provide richer and greater insight into students" experiences with and uptake of hybrid feedback.

5.3 Implications

Informed by the findings of the study, theoretical and practical recommendations are presented below. The practical implications are relevant to two broad types of stakeholders: those who would be using the tools (e.g., program directors, curriculum writers and students) and AWE developers. The chapter concludes with implications for future research.

5.3.1 Implications for Theory

The study found a lack of variance in scores from the AWE system compared to rater scores, suggesting that Criterion may examine somewhat different dimensions of writing than human raters. AWE systems' scoring measures approximations of writing quality because the scoring systems are trained to predict human ratings by estimating trins with proxes. Trins are intrinsic characteristics of writing, whereas proxes are approximations of those characteristics such as length of a word may be a proxy for lexical sophistication (Page & Petersen, 1995). Also, different AWE systems' scoring engines use different weights and proxes for different writing dimensions. Therefore, researchers agree that while AWE scores can predict human scores, measuring the same construct requires more than predictive accuracy (Deane et al., 2013). In addition, due to the 'black box' nature of commercially available AES system, "we cannot know the criteria by which the computer scores the writing and so we cannot understand the kinds of bias that may have been built into the scoring" (Attali, 2013, p. 18). However, this line of research is impossible until developers of AWE systems share their algorithms.

The study has provided support for the hypothesis that AWCF combined with teacher feedback has the potential to facilitate writing development. The study's findings have given some insights into how and why writing development occurs due to hybrid feedback. As with previous studies on WCF, students found that automated feedback helped them to notice errors, and the metalinguistic feedback helped them resolve the errors (Bitchener & Storch, 2016). The findings suggest that hybrid feedback may have helped the students develop the skills to become more autonomous in their learning. We need to be cautious, however, as the use of the new technology may lead to losses. For example, when a new tool helps students by giving them a new word, or a correction, out of context, students might lose autonomy and skills such as searching, figuring out, or learning by other means. Additionally, students reported that they preferred the targeted approach, where the feedback is tailored to students" needs, especially for lower proficiency students, because it reduced their cognitive load and may facilitate the attention required to process new information.

Due to their limitation, AWE systems currently only give unfocused or comprehensive written feedback that is abstract and divorced from context and which does not include salient examples or alternative models. Little research has investigated the effectiveness of unfocused written corrective feedback, and the findings are mixed. However, the research seems to suggest that advanced learners at can better attend to unfocussed feedback because they are able to process a wider range of input in a single feedback session (Van Beuningen et al., 2012). The findings in this study seem to corroborate these findings: students with lower in proficiency were less satisfied with the unfocused nature of AWCF, which may be due to the fact that the feedback was in the target language and the metalinguistic feedback was abstract and not specific. Furthermore, previous research found that both direct and indirect WCF had a positive effect, but the awareness the learners develop varies depending on the type and degree of metalinguistic explanation they receive (Stefanou & Révész, 2015). Therefore, future research should examine which combinations of types of feedback and metalinguistic explanations facilitate more awareness. As noted before, the two types of metalinguistic explanations AWE systems provide are generic and specific, with generic feedback consisting of a canned response that offers no specific remedies while specific feedback incorporating some text component to give a recommendation. Future research can employ an experimental design with different groups of learners receiving different types of feedback and explanations. Such research can employ stimulated recall to investigate students' output and their writing processes in response to the different types of feedback.

The study also provides support for SCT views of learning and the importance SCT attributes to human interaction. The findings suggest that students engaged more with teacher feedback that is graduated, contingent, and dialogic; receiving teacher feedback after AWE feedback helped to mediate machine feedback by considering the needs and proficiencies of the learners as recommended by researchers who have called for a hybrid approach (Attali et al., 2013; Mehrabi-Yazdi, 2018). In addition, although generic, AWE feedback on successive drafts could be perceived as a form of dialogue because the

feedback changes due to revisions made by the student (Bitchener & Storch, 2016). In the hybrid approach to feedback, AWE as a material tool not only enables actions to take place, such as enabling students to notice and resolve the feedback but also shapes actions. That is, AWE can shape students' notions of what good writing is, reinforce iterative aspects of writing (Nassaji & Swain, 2000) and delimit the forms and genres of possible writing projects, including multimodal composition and collaborative authorship (Thumlert et al., 2015). Although studies on hybrid feedback are nascent, more research is needed on how AWE as a mediation tool may promote or limit writing development and performance. As Hirvela et al. (2016) commented, "students not only need help in how to compose, but also in understanding how texts are shaped by topic, audience, purpose and cultural norms so they can activate schemata, genre awareness, grammar proofing, and responsiveness to a particular audience" (p. 48). However, it is unclear how integrating AWE in the classroom shapes or limits students' writing views and writing development.

5.3.2 Implications for Integration, Instruction, and Implementation

A key finding of this study is that hybrid feedback can facilitate the improvement of EAP students' writing by fostering the revision process. While the hybrid feedback helped students adopt a more process-oriented approach to writing, the extent of this effect varied across individual students, suggesting that different learners benefit differently from AWC feedback depending on such factors as motivation, learning style, and learning goals, which emphasizes the role of teachers as facilitators between AWE and students. Therefore, writing teachers should take into consideration how AWE is integrated into the classroom. It is suggested that teachers pay more attention to the social and communicative aspects of writing when an AES system, such as Criterion, is integrated into the classroom (Wang & Brown, 2008; Ware & Warschauer, 2006) . Chen and Cheng (2008) noted that students favour a feedback process that is social and communicative. However, as reflected in students' comments in this study, AES systems fail to fill the social gap of meaning negotiation because most AES systems "are

theoretically grounded in a cognitive information-processing model, which does not focus on the social and communicative dimensions of writing" Chen and Cheng (2008, p. 9). Therefore, when integrating AWE in the classroom, teachers need to consider the importance of meaningful communication between the writer and the reader by focusing on content, development, and organization so that students recognize that a human response in such matters is not just a pattern-matching algorithm.

When students in this study did not resolve AWE feedback, there were three broad reasons for this as seen from the individual case analyses: insufficient metalinguistic feedback, not trusting the feedback, and choosing which feedback to attend to because of the comprehensive nature of AWE feedback. Each reason requires teacher facilitation. First, a central theme in the focus-group data was that students did not understand Criterion's generic and vague feedback; therefore, they focused on more concrete feedback related to surface-level errors. In addition, because the effects of different errors vary across contexts (Ranalli, 2018), students need to be made aware of which errors affect the message's communicability and focus on and resolve those errors. Therefore, it is suggested that teachers support AWE feedback by teaching the terms that AWE systems use because there are many synonyms for the same structures, such as present progressive and present continuous. In other words, teachers need to mediate AWE feedback and ensure that students have the capacity to understand the metalinguistic feedback of AWE. Teachers also need to support students by helping them decide when to resolve or disregard Criterion feedback. For instance, some students in this study did not understand when to use passive and active voices in writing. Teachers could explain to students that flagging of the passive voice by Criterion does not necessarily mean that it needs to be corrected and remind them that it is essential to use passive structures in writing when who did the action is not important or known, although it may make their sentences wordy and indirect. In other words, teachers can have a lesson to determine if AWE feedback on the use of passive voice needs to be resolved, how to resolve such feedback, and when to ignore the feedback. Also, due to the comprehensive nature of Criterion feedback, teachers should help students be cognizant of the most critical and

vital errors to resolve. For example, teachers can help individual students focus on specific errors that each student needs to work on.

Second, both teachers and students need to be aware of AWE feedback's limitations. Teachers need to understand the shortcomings of AWE systems to help address students' doubts when they use AWE feedback and give concrete advice on improving students' writing. Likewise, with additional professional development related to AWE that stresses the importance of adopting one"s classroom instruction when AWE systems are integrated, teachers would be able to find more efficient ways to adapt AWE features for their classroom context. Allocating sufficient time and support before integration to better understand AWE feedback's affordances and limitations is necessary. Researchers like Ware (2011) have noted that teachers who only have minimal training on the software might not be aware of the tool's versatilities and options. They may be dismissive of AWE feedback before critically engaging with the software.

In addition, students need to be aware of AWE feedback's limitations because AWE feedback is not always accurate and can generate false positives and incorrect error codes. If students take AWE feedback at face value and are not cognizant of its limitations, they may not trust the feedback or engage with it (Zhang, 2020). For example, the findings of this study show that some students expressed doubts about some of the error codes they received from the AWE system, while others were confident about false positives as in previous research (Zhang, 2017). However, inaccurate WCF from AWE systems may still be helpful for learners when they are aware of these limitations and taught how to evaluate and use such feedback in conjunction with the teachers' mediation of AWE feedback. WCF from AWE systems can encourage learners to be more cognizant of the writing and revision processes if they receive training on evaluating and using such feedback (Lavolette et al., 2015). Grimes and Warschauer (2010) posited that erroneous feedback is most problematic when presented as authoritative, and there is no human intervention to override the feedback. In cases of inaccurate feedback and error codes from AWE systems, the teacher can help remove self-doubts and help students

move forward with appropriate strategies to double-check errors and the feedback they receive.

Third, although both groups in the current study have relatively high proficiency (overall IELTS 5.5 indicating upper intermediate proficiency) in writing, some of them still found it frustrating when they received generic feedback and could not resolve it. Previous research on WCF found greater potential for focused feedback because unfocused WCF may impose a cognitive overload on the learner (Wang & Jiang, 2015). Contrary to studies on focused feedback effectiveness, Criterion provides comprehensive feedback due to technological limitations. Cognitive overload is a concern not only for lower proficiency learners but also for higher proficiency students. For example, one of the students in this study may have engaged superficially with the feedback because they were focused on the quantity ("hunting down all errors") rather than the impact or quality of errors. Therefore, teachers can address student-specific issues by devoting time during oral feedback sessions or in-class instruction to discussing error types and helping students focus on the most relevant error types. This mediation may help to mediate the generalized nature of AWE feedback and reduce cognitive overload.

In addition, although the classroom context in the current study did not involve high-stakes testing, the class was high-stakes because it acted as a gate for students entering their undergraduate programs at the university. Due to the nature of the program and that the in-class writing tasks would directly affect students' grades, there may inevitably be some possible washback effects of the AWE system on the way students write and compose essays. In other words, the students may intentionally or unintentionally write in accordance with the parameters established by Criterion. However, unlike other studies that used Criterion scores as part of students' grades, this study ultimately only used teacher ratings for grading. Although the study did not examine the washback effects of the use of Criterion on students' writing, some students noted a drive to reduce all mistakes identified by Criterion in their writing. This may lead to students attempting to "please" AES systems and paying more attention to form than content, as found in

previous research (Huang, 2014). However, the urge to revise and correct may have led to positive washback as suggested by the Writing Process Questionnaire, which showed that students tended to reflect more on their writing processes, to think more about how their ideas relate to each other, to organize their ideas in a more logical order, and to engage in more monitoring and revising. Therefore, teachers should consider how AES is implemented in the classroom to reduce potential adverse washback effects. For example, teachers should reduce situations where students focus on getting a high score from the machine and use it primarily as a means of giving feedback for formative purposes (Woodworth & Barkaoui, 2020). In addition, if the goal of the class is process-oriented, the hybrid approach to feedback may induce positive washback effects in the form of constant and immediate feedback and facilitation of opportunities for in-class and independent writing practice and revisions (Huang, 2014).

5.3.3 Implications for AWE System Developers

The findings of this research have four general implications for developers of AWE systems. First, a critical concern is that the students in this study were not satisfied with the Criterion holistic and trait scores. In the focus group, students revealed that they were frustrated with the scores not changing even when all the errors were resolved. Students commented that no matter how much vocabulary they changed, their score for word choice remained the same. Even with no error messages, the students may not get the highest score for grammar, usage, mechanics, and conventions. Perhaps more details can be provided with the scores by the AWE system to explain to the students why they have received lower than perfect scores when all feedback for the trait has been resolved.

Second, during the focus group, students reiterated that they wanted more specific feedback that catered to their proficiency level. Therefore, AWE systems need to be more adaptable and flexible to meet the needs of students with varying proficiency levels. Research in the field also suggests that feedback should be adaptable and flexible. As Criterion does not give feedback based on the students' proficiency level, it should be more versatile by allowing end-users to disable certain feedback functions or have a set

of graduated feedback functions enabled or disabled after a self-guided questionnaire to match individual learners" preferences and stages of L2 development. The uptake behaviour of some students for revisions shows that there was only superficial engagement with Criterion feedback for word choice and grammar apart from punctuation and spelling. This may indicate that generic feedback for case-specific error types may limit the comprehensibility of the metalinguistic explanations from Criterion. Accordingly, this may suggest a lack of meaningful engagement when seeking elaboration and clarification for the feedback received. To help provide feedback that is more graduated, contingent, and dialogic (Bitchener & Storch, 2016), Criterion could allow learners to filter the feedback to the level of students' proficiency to provide more contingent and graduated feedback. This was previously suggested in research for a function to toggle particular error types (Ranalli, 2018), which would lead to a more selective approach to feedback (Mehrabi-Yazdi, 2018), resulting in reduced cognitive overload. For example, a teacher (or the student) could opt to only highlight feedback for verb tense for a particular student to focus their attention on that specific error type. This, in turn, would give the teacher more flexibility to adjust the focus of the feedback to align with the curricular goals of the class and the proficiency level, or the current metalinguistic knowledge, of the students (Ranalli, 2018). A similar approach could be taken for other types of errors or feedback.

Likewise, an option to give only specific or generic feedback would be helpful in the classroom. In this way, the teacher can better support the students' needs and abilities and the writing task's instructional focus. For example, a teacher can toggle on a specific grammar point that a student needs help with to make AWE feedback more contingent, graduated, and personalized to the needs of individual students because, as it stands, the technical capacities of AWE systems are too limited to give individualized feedback. Analysis of AWE feedback by Mehrabi-Yazdi (2018) found that such feedback lacks coherence because it is neither staged nor purposeful and "not regulated and delivered with regard to what feedback has already been provided to the student in previous submissions" (p. 904).

Third, although the focus of this study was on three types of writing tasks, the class included other writing assignments such as reading summaries, reflections, and a research paper. However, Criterion does not distinguish different genres and writing tasks, which could be added in future revisions. A common theme during the focus group was that the students were frustrated with the repeated highlights, repeated words, and the construction of passive sentences. Research has shown that the increase of passive forms may indicate higher proficiency in academic writing (see Biber & Gray, 2010) and treating passive structures similarly in an academic essay as in other non-academic genres shows that Criterion is insensitive to the dynamic nature of lexico-grammatical choices for different purposes and genres of writing. This calls for more flexibility of AWE for other genres and forms of writing. Perhaps the end-user can select the genre and tone of writing as they do in newer grammar checkers such as Grammarly and ProWritingAid.

Lastly, multiple students reported in the interview that hunting for errors was similar to playing a video game, which motivated them to resolve as many mistakes as possible. Consequently, to enhance the integration of AWE systems into the class, including gamified mechanics such as points, badges, and leaderboards in AWE systems can motivate the students to revise and monitor their writing more often and at multiple levels. Also, students found that as they became used to the tool, they were more efficient at fixing the errors; yet they found the onboarding process to using the tool confusing. A game-like onboarding process may be a natural extension to leverage a smoother onboarding process to facilitate more familiarity with AWE systems. Gamified elements in education have been shown to increase motivation and engagement (Hanus & Fox, 2015). Student motivation is a complex dynamic system and gamified elements can be part of that system to interact with the intentionality of the technologically mediated world. Teachers see an increasing number of L2 learners who are part of a generation that Prensky (2001) has described as "Digital Natives." The gamification of AWE systems and the addition of game-like features may enhance students' motivation to write more frequently and seek and use feedback to improve their L2 writing.

5.3.4 Implications for Research

This study has shown that a hybrid approach to feedback can effectively be used in the classroom. However, further research is needed on how AWE can be integrated into the classroom effectively and efficiently. In addition, understanding the delicate balance between the two types of feedback and how students use the combination is necessary to implement hybrid feedback fully.

Learners' uptake and retention of hybrid feedback can be studied over more extended periods with a larger sample of students and a larger number of writing tasks to identify richer evidence of change. Furthermore, future studies need to be conducted in different programs with a diverse population to determine how hybrid feedback could be applied to learners with varying proficiency levels, different cultural and linguistic backgrounds, and course curricula. Also, with a larger number of writing samples from a greater randomized pool of learners, there would be less chance of outliers and confounding variables to impact the analysis, and the findings would be more generalizable. In addition, to triangulate self-reported data, concurrent protocols that do not cause greater cognitive load, such as keystroke logging, eye tracking, and recording, in addition to retrospective protocols and individualized interviews, can provide a fuller picture of learners' use and views of hybrid feedback.

Though the outcomes of this study are positive, more research needs to be done to examine how hybrid feedback can be integrated into naturalistic writing classes for optimal effectiveness. Like WCF research that employs different feedback types, future studies should ask what combination of teacher and machine feedback would be most beneficial for different learners. Previous research recommends some labour division between the teacher and the AWE system in terms of feedback, where the teacher focuses on higher-level wiring skills (Chen & Cheng, 2008; Li et al., 2015; Stevenson, 2016). The findings of this study show that the division between teachers and AWE systems is not so clear. Future studies on hybrid feedback can investigate different types of combinations

of feedback with specific learner populations to provide empirical evidence that informs the pedagogical implementation of hybrid feedback systems for different learners and contexts. For example, learners can be placed in a different group with each group receiving a different combination of automated and teacher feedback: all feedback from AWE and all feedback from a teacher, form-focused feedback from AWE and content-related feedback from the teacher, etc. Also, research on how different teachers mediate AWE feedback can reveal more nuanced information about ways to implement hybrid feedback in different contexts. For example, in a study by Chen and Cheng (2008), the authors gave access to AWE software to three different teachers where each teacher utilized the machine differently with different levels of success. Therefore, more research is needed to inform teachers of the most effective ways to integrate AWE into their writing instruction contexts.

In addition, like previous research on focused and unfocused WCF, the efficacy of generic and specific AWE feedback could be examined to gain a deeper understanding of best practices when using hybrid feedback in the classroom. For example, previous research has shown that different types of feedback have different effects on learners' depth of processing (Kim & Bowles, 2019; Ling et al., 2021). Similarly, teacher and automated feedback may have different effects on different aspects of L2 writing development. Likewise, more research is needed to understand how teachers should change and adapt their feedback when combined with AWE feedback. While it has been shown that AWE can save time for teachers by reducing or omitting some surface-level feedback, future studies should examine how teachers can give feedback on content and organization to integrate automated feedback in the classroom better.

Moreover, future studies need to consider the effects of individual learner differences such as attitudes to AWE and teacher feedback, the importance of English writing in their future, L1, and L2 proficiency level, on the effectiveness of hybrid feedback. In particular, a deeper understanding of learners' cognitive and affective engagement with hybrid feedback and its effects on L2 writers' revision strategies and L2 acquisition is

needed. For example, researchers can use a case-study approach and/or an experimental design to compare how learners with different profiles engage with hybrid feedback. This would provide a more complete picture of learners' engagement with hybrid feedback and allow teachers to adapt hybrid feedback to learners' needs and characteristics. To achieve this, future research can employ quasi-experimental mixed methods design to give voice to participants to ensure that study findings are grounded in participants' experiences because how learners view hybrid feedback is an essential factor in their engagement with it.

Lastly, much of the field has focused on investigating the efficacy of AWE in increasing writing quality. However, very little has been done to investigate how learners process AWE feedback as socially mediated actions in culture-specific contexts. For instance, there is a concern that AWE feedback is modelled after very strict and limited types of essays in specific contexts. The type of feedback the AWE system provides may alter learners' perception of what good writing is. As noted earlier, learners' perceptions of good writing would be changed if their audience is a machine, and the types of writing programmed into the AWE system may limit what types of writing would be considered necessary in the classroom. Therefore, there is a need to examine how teacher feedback can be combined to mitigate these potential adverse washback effects of the use of AWE systems in the classroom and avoid limiting learners' perception of writing. This can be achieved by using a case-study approach to examine how learners process AWE feedback for different types of essays, including writing genres that are more diverse, with different types of teacher mediations. It is hoped that such research will improve our understanding of the effectiveness of hybrid feedback in different contexts and inform L2 writing instruction.

References

Adams, K. A., & Lawrence, E. K. (2018). *Research methods, statistics, and applications*. Sage Publications.

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.

Alshahrani, A., & Storch, N. (2014). Investigating teachers" written corrective feedback practices in a Saudi EFL context: How do they align with their beliefs, institutional guidelines, and students" preferences. *Australian Review of Applied Linguistics*, *37(2)*, 101-122.

Andrade, H., Huff, K., & Brooke, G. (2012). Assessing learning. *Education Digest*, *78(3)*, 46-53.

Anson, C. M. (2006). Can"t touch this: Reflections on the servitude of computers as readers. In P. F. H. Ericsson, R (Ed.), *Machine scoring of human essays* (pp. 38-56). Logan, UT: Utah State University Press.

Ashwell, T. (2000). Patterns of teacher response to student writing in a multiple-draft composition classroom: Is content feedback followed by form feedback the best method. *Journal of second language writing*, *9(3)*, 227-257.

Attali, Y. (2013). Validity and reliability of automated essay scoring. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 181-198). Routledge.

Attali, Y. (2015). Reliability-based feature weighting for automated essay scoring. *Applied psychological measurement*, *39(4*, 303-313.

Attali, Y. (2004). Exploring the feedback and revision features of Criterion. *Journal of Second Language Writing*, *14*, 191-205.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with E-Rater® V.2.0. *ETS Research Report Series*, *4*, i-21.

Attali, Y., Lewis, W., & Steier, M. (2013). Scoring with the computer: Alternative

procedures for improving the reliability of holistic essay scoring. *Language Testing*, *30*, 125-141.

Bai, L., & Hu, G. (2017). In the face of fallible AWE feedback: how do students respond. *Educational Psychology*, *37(1)*, 67-81.

Balota, D. A., & Chumbley, J. I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human perception and performance*, *10(3)*, 340.

Barkaoui, K., & Hadidi, A. (2020). *Assessing Change in English Second Language Writing Performance*. Routledge.

Beers, S. F., & Nagy, W. E. (2009). Syntactic complexity as a predictor of adolescent writing quality: Which measures? Which genre. *Reading and Writing*, *22(2)*, 185-200.

Bertelsen, O. W., & Bødker, S. (2003). Activity theory. In J. M. Carroll (Ed.), *HCI models, theories, and frameworks: Toward a multidisciplinary science* (pp. 291-324). San Francisco: Morgan Kaufmann.

Bestgen, Y., & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing*, *26*, 28-41.

Biber, D., Gray, B., Staples, S., & Egbert, J. (2020). Investigating grammatical complexity in L2 English writing research: Linguistic description versus predictive measurement. *Journal of English for Academic Purposes*, *46*, 1-15.

Biber, D., & Gray, B. (2010). Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes*, *9(1)*, 2-20.

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly*, *45*, 5-35.

Bitchener, J. (2008). Evidence in support of written corrective feedback. *Journal of Second Language Writing*, *17*, 102-118.

Bitchener, J. (2012). A reflection on "the language learning potential" of written CF.

Journal of Second Language Writing, *21*, 348-363.

Bitchener, J., & Ferris, D. R. (2012). *Written corrective feedback in second language acquisition and writing*. Routledge.

Bitchener, J., & Knoch, U. (2008). The value of written corrective feedback for migrant and international students. *Language Teaching Research*, *12*, 409-431.

Bitchener, J., & Knoch, U. (2009a). The contribution of written corrective feedback to language development: A ten month investigation. *Applied linguistics*,

Bitchener, J., & Knoch, U. (2009b). The relative effectiveness of different types of direct written corrective feedback. *System*, *37*, 322-329.

Bitchener, J., & Knoch, U. (2010). Raising the linguistic accuracy level of advanced L2 writers with written corrective feedback. *Journal of Second Language Writing*, *19*, 207-217.

Bitchener, J., & Knoch, U. (2015). Written corrective feedback studies: Approximate replication of Bitchener & Knoch (2010a) and Van Beuningen, de Jong & Kuiken (2012). *Language Teaching*, *48*, 405-414.

Bitchener, J., & Storch, N. (2016). *Written corrective feedback for L2 development*. Multilingual Matters Limited.

Bitchener, J., Young, S., & Cameron, D. (2005). The effect of different types of corrective feedback on ESL student writing. *Journal of Second Language Writing*, *14*, 191-205.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, *3(2)*, 77-101.

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

Brown, H. D. (2000). *Principles of language learning and teaching* (4). New York: Longman.

Bulté, B., & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing*, *26*, 42-65.

Burstein, J., Chodorow, M., & Leacock, C. (2003, August). *Criterion online essay*

evaluation: An application for automated evaluation of student essays. In Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence, Acapulco, Mexico.

Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The Criterion online writing service. *AI Magazine*, *25(3)*, 27-36.

Calfee, R. C., Miller, R. G., & Graham…, S. (2007). Best practices in writing assessment. *... practices in writing ...*,

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied linguistics*, *1*, 1-47.

Casal, J. E., & Lee, J. J. (2019). Syntactic complexity and writing quality in assessed first-year L2 writing. *Journal of Second Language Writing*, *44*, 51-62.

Chan, S., Bax, S., & Weir, C. (2017). Researching participants taking IELTS Academic Writing Task 2 (AWT2) in paper mode and in computer mode in terms of score equivalence, cognitive validity and other factors. *IELTS Research Reports Online Series*, 47.

Chan, S. H. C. (2018). Some evidence of the development of L2 reading-into-writing skills at three levels. *Language, Education and Assessment*,

Chandler, J. (2000). The efficacy of error correction for improvement in the accuracy of L2 student writing. *AAAL Conference*,

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*, 267-296.

Chapelle, C. A., Chung, Y.-R., & Xu, J. (2008). Introduction. In C. A. Chapelle, Y.-R. Chung, & J. Xu (Eds.), *Towards adaptive CALL natural language processing for diagnostic language assessment* (pp. 1-7). Iowa State University.

Chapelle, C. A., & Sauro, S. (2017). *The handbook of technology and second language teaching and learning* (1 ed.). Wiley-Blackwell.

Chen, C.-F. E., & Cheng, W.-Y. E. C. (2008). Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes. *Language Learning & Technology*, *12(2)*, 94-112.

Chen, J., Zhang, M., & Bejar, I. I. (2017). An Investigation of the e-rater ® Automated Scoring Engine"s Grammar, Usage, Mechanics, and Style Microfeatures and Their Aggregation Model. *ETS Research Report Series*, *2017*, 1-14.

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

Chiang, S. (2003). The importance of cohesive conditions to perceptions of writing quality at the early stages of foreign language learning. *System*, *31(4)*, 471-484.

Choi, I. C. (2008). The impact of EFL testing on EFL education in Korea. *Language Testing*, *25*, 39-62.

Cohen, A. D. (1987). Student processing of feedback on their compositions. In A. L. Wenden & J. Rubin (Eds.), *Learner strategies in language learning* (pp. 57-69). Englewood Cliffs, NJ: Prentice-Hall.

Cohen, A. D., & Cavalcanti, M. C. (1990). Feedback on compositions: Teacher and student verbal reports. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 155-177). Cambridge: Cambridge University Press.

Cohen, A. D., & Robbins, M. (1976). Toward assessing interlanguage performance: The relationship between selected errors, learners" characteristics, and learners" explanations. *Language learning*, *26(1)*, 45-66.

Communication, Conference on College Composition and. (2004). CCCC position statement on teaching, learning, and assessing writing in digital environments. https://cccc.ncte.org/cccc/resources/positions/digitalenvironments.

Connor, U., & Mbaye, A. (2002). Discourse approaches to writing assessment. *Annual Review of Applied Linguistics*, *22*, 263-278.

Creswell, J. W., & Clark, V. L. P. (2007). *Designing and conducting mixed methods research.* Thousand Oaks, CA: Sage publications.

Cross, R., & O"Loughlin, K. (2013). Continuous assessment frameworks within university English Pathway Programs: realizing formative assessment within high-stakes contexts. *Studies in Higher Education*, *38(4)*, 584-594.

Crossley, S., & McNamara, D. (2011). *Text coherence and judgments of essay quality:*

Models of quality and coherence. In Proceedings of the Annual Meeting of the Cognitive Science Society, 33(33) (pp. 1236-1241).

Crossley, S. A., Bradfield, F., & Bustamante, A. (2019a). Using human judgments to examine the validity of automated grammar, syntax, and mechanical errors in writing. *Journal of Writing Research*, *11(2)*, 251-270.

Crossley, S. A., Kyle, K., & Dascalu, M. (2019b). The Tool for the Automatic Analysis of Cohesion 2.0: Integrating semantic similarity and text overlap. *Behaviour research methods*, *51(1)*, 14-27.

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, *35(2)*, 115-135.

Crossley, S. A., & Salsbury, T. (2010). Using lexical indices to predict produced and not produced words in second language learners. *The Mental Lexicon*, *5(1)*, 115-147.

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011). Predicting lexical proficiency in language learner texts using computational indices. *Language Testing*, *28(4)*, 561-580.

Crossley, S. A., Subtirelu, N., & Salsbury, T. (2013). Frequency effects or context effects in second language word learning: What predicts early lexical production. *Studies in Second Language Acquisition*, *35(4)*, 727-755.

Crossley, S. A., Allen, L. K., Snow, E. L., & McNamara, D. S. (2015a). *Pssst. textual features. there is more to automatic essay scoring than just you!* In Proceedings of the Fifth International Conference on Learning Analytics And Knowledge - LAK "15, (pp. 203-207).

Crossley, S. A., Kyle, K., Allen, L. K., Guo, L., & McNamara, D. S. (2014). Linguistic microfeatures to predict L2 writing proficiency: A case study in Automated Writing Evaluation. *The Journal of Writing Assessment*, *7*, 1-16.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing*, *32*, 1-16.

Crossley, S. A., Kyle, K., & Mcnamara, D. S. (2015b). To aggregate or not ? Linguistic

features in automatic essay scoring and feedback systems. *The Journal of Writing Assessment*, *8*, 1-14.

Crossley, S. A., & McNamara, D. S. (2016). Say more and be more coherent: How text elaboration and cohesion can increase writing quality. *Journal of Writing Research*, *7(3)*, 351-370.

Cumming, A. (2015). Theoretical orientations to L2 writing. In *Handbook of second and foreign language writing* (pp. 65-90). Walter de Gruyter Boston, MA.

Cumming, A., Kantor, R., Baba, K., Eouanzoui, K., Erdosy, U., & Jamse, M. (2006). Analysis of discourse features and verification of scoring levels for independent and integrated prototype written tasks for the new TOEFL. *ETS Research Report Series*, *2006(1)*, i-77.

Cummins, J., & Davison, C. (2007). *International handbook of English language teaching*. Springer.

Dann, R. (2002). *Promoting assessment as learning: Improving the learning process*. Routledge.

Davidson, F. (1991). Statistical support for training in ESL composition rating. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 155-164). Ablex Publishing Corporation.

Davies, M. (2013). Google Scholar and COCA-Academic: Two very different approaches to examining academic English. *Journal of English for Academic Purposes*, *12(3)*, 155-165.

De Bot, K. (1996). The psycholinguistics of the output hypothesis. *Language learning*, *46(3)*, 529-555.

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, *18(1)*, 7-24.

Deane, P., Williams, F., Weng, V., & Trapani, C. S. (2013). Automated essay scoring in innovative assessments of writing from sources. *The Journal of Writing Assessment*, *6(1)*, 40-56.

Deane, P., & Zhang, M. (2015). Exploring the feasibility of using writing process features to assess text production skills. *ETS Research Report Series*, *2015*, 1-16.

Deci, E., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behaviour*. Springer Science & Business Media.

DeKeyser, R. (2007). *Practice in a second language: Perspectives from applied linguistics and cognitive psychology*. Cambridge University Press.

Denzin, N. K., & Lincoln, Y. S. (1994). *Handbook of qualitative research.* Sage publications, inc.

Deterding, S., Sicart, M., Nacke, L., O"Hara, K., & Dixon, D. (2011, May 7-12). *Gamification. using game-design elements in non-gaming contexts.* In CHI"11 extended abstracts on human factors in computing systems, (pp. 2425-2428).

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.

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, *5(1)*,

Duijnhouwer, H., Prins, F. J., & Stokking, K. M. (2012). Feedback providing improvement strategies and reflection on feedback use: Effects on students" writing motivation, process, and performance. *Learning and Instruction*, *22(3)*, 171-184.

El Ebyary, K., & Windeatt, S. (2010). The impact of computer-based feedback on students" written work. *International Journal of English Studies*, *10(2)*, 121-142.

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in second language acquisition*, 24*(2)*, 143-188.

Ellis, R. (2010a). Cognitive, social, and psychological dimensions corrective feedback. In R. Batstone (Ed.), *Sociocognitive Perspectives on Language Use and Language Learning* (pp. 1 51-1165). OUP Oxford.

Ellis, R. (1997). *Second Language Acquisition*. Oxford University Press.

Ellis, R. (2009). A typology of written corrective feedback types. *ELT Journal*, *63*, 97-107.

Ellis, R. (2010b). Epilogue: A framework for investigating oral and written corrective feedback. *Studies in Second Language Acquisition*, *32*, 335-349.

Ellis, R., & Barkhuizen, G. (2005). *Analysing Learner Language*. Oxford University Press.

Ellis, R., Sheen, Y., Murakami, M., & Takashima, H. (2008). The effects of focused and unfocused written corrective feedback in an English as a foreign language context. *System*, *36*, 353-371.

Erdosy, M. U. (2004). Exploring variability in judging writing ability in a second language: A study of four experienced raters of ESL compositions. *ETS Research Report Series*, *2004(1)*, i-62.

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

Erlam, R., Ellis, R., & Batstone, R. (2013). Oral corrective feedback on L2 writing: Two approaches compared. *System*, *41(2)*, 257-268.

Evans, N. W., Hartshorn, K. J., & Tuioti, E. A. (2010). Written corrective feedback: Practitioners" perspectives. *International Journal of English Studies*, *10*, 47-77.

Evans, S., & Green, C. (2007). Why EAP is necessary: A survey of Hong Kong tertiary students. *Journal of English for Academic Purposes*, *6(1)*, 3-17.

Evans, S., & Morrison, B. (2010). The first term at university: Implications for EAP. *ELT journal*, *65(4)*, 387-397.

Faigley, L., & Witte, S. (1981). Analyzing revision. *College composition and communication*, *32(4)*, 400-414.

Fathman, A. (1990). Teacher response to student writing: Focus on form versus content. *Second language writing: Research insights for the classroom*, 178-190.

Fazio, L. L. (2001). The effect of corrections and commentaries on the journal writing accuracy of minority-and majority-language students. *Journal of second language writing*,

Feng, H.-H., Saricaoglu, A., & Chukharev-Hudilainen, E. (2015). Automated error detection for developing grammar proficiency of ESL learners. *CALICO Journal*, *33(1)*, 49-70.

Ferris, D. R. (1995a). Can advanced ESL students be taught to correct their errors?

Ferris, D. R. (1999). The case for grammar correction in L2 writing classes: A response to

Truscott (1996). *Journal of second language writing*, *8(1)*, 1-11.

Ferris, D. R. (2002). Teaching students to self-edit. In J. C. Richards & W. A. Renandya (Eds.), *Teaching students to self-edit. Methodology in language teaching: An anthology of current practice* (pp. 328-334).

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

Ferris, D. R. (1997). The influence of teacher commentary on student revision. *Tesol Quarterly*, *31(2)*, 315-339.

Ferris, D. R. (2003). *Response to student writing: Implications for second language students*. Routledge.

Ferris, D. R. (1994). Lexical and syntactic features of ESL writing by students at different levels of L2 proficiency. *TESOL Quarterly*, *28*, 414.

Ferris, D. R. (1995b). Teaching students to self-edit. *TESOL Journal*, *4*, 18-22.

Ferris, D. R. (2004). The "grammar correction" debate in l2 writing: where are we, and where do we go from here? (and what do we do in the meantime.?). *Journal of Second Language Writing*, *13*, 49-62.

Ferris, D. R. (2006). Does error feedback help student writers? New evidence on the short-and long-term effects of written error correction. In K. Hyland & F. Hyland (Eds.), *Feedback in second language writing: Contexts and Issues* (pp. 81-147). Mahwah, NJ: Lawrence Erlbaum Associates.

Ferris, D. R. (2010). Second language writing research and written corrective feedback in SLA: Intersections and practical applications. *Studies in Second Language Acquisition*, *32*, 181-201.

Ferris, D. R. (2014). Responding to student writing: Teachers" philosophies and practices. *Assessing Writing*, *19*, 6-23.

Ferris, D. R., Chaney, S. J., Komura, K., Roberts, B. J., & McKee, S. (2000). Perspectives, problems, and practices in treating written error. In *Colloquium presented at International TESOL Convention*, Vancouver, B.C., (pp. 14-18).

Ferris, D. R., & Hedgcock, J. S. (2013). *Teaching L2 composition: Purpose, process, and practice*. Routledge.

Ferris, D. R., John S. Hedgcock, John S. (2013). *Teaching L2 composition: Purpose, process, and practice*. Routledge.

Ferris, D. R., Liu, H., Senna, M., & Sinha, A. (2010). *Written corrective feedback and individual variation in L2 writing.* In.

Feuerstein, R., Rand, Y., Hoffman, M., & Miller, R. (1980). *Instructional enrichment*. Baltimore: University Park Press.

Feuerstein, R., Rand, Y., & Hoffman, M. B. (1979). *The dynamic assessment of retarded performers: The learning potential assessment device, theory, instruments, and techniques.* Glenview: Scott Foresman & Co.

Feuerstein, R., Rand, Y., & Rynders, J. E. (1988). The learning potential assessment device. In *Don"t Accept Me as I am* (pp. 191-207). Boston, MA: Springer.

Ferris, D. R. & Roberts, B. (2001). Error feedback in L2 writing classes: How explicit does it need to be. *Journal of second language writing*, 10(3), 161-184.

Feuerstein, T. (1990). The theory of structural cognitive modifiability. *DOCUMENT RESUME*,

Field, A. (2018). *Discovering Statistics Using IBM SPSS Statistics*. SAGE Publications Limited.

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College composition and communication*, *32(4)*, 365-387.

Fukuta, J., Tamura, Y., & Kawaguchi, Y. (2019). Written languaging with indirect feedback in writing revision: is feedback always effective. *Language awareness*, *28(1)*, 1-14.

Gass, S. M. (1997). *Input, interaction, and the second language learner*. Lawrence Erlbaum Associates Publishers.

Ghufron, M. A. (2019, April). *Exploring an automated feedback program "grammarly" and teacher corrective feedback in EFL writing assessment: Modern vs. traditional assessment.* In Proceedings of the 3rd English Language and Literature International Conference.

Griffiths, L., & Nicolls, B. (2010). e-Support4U: An evaluation of academic writing skills

support in practice. *Nurse Education in Practice*, *10(6)*, 341-348.

Grimes, D., & Warschauer, M. (2010). Utility in a fallible tool: A multi-site case study of automated writing evaluation. *The Journal of Technology, Learning and Assessment*, *8(6)*,

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

Guénette, D. (2007). Is feedback pedagogically correct? Research design issues in studies of feedback on writing. *Journal of Second Language Writing*, *16*, 40-53.

Guichon, N., & Cohen, C. (2016). Multimodality and CALL. *hal.archives-ouvertes.fr*,

Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing*, *18*, 218-238.

Halliday, M. A. K., & Hasan, R. (2014). *Cohesion in English*. Routledge.

Han, Y., & Hyland, F. (2015). Exploring learner engagement with written corrective feedback in a Chinese tertiary EFL classroom. *Journal of Second Language Writing*, *30*, 31-44.

Hanus, M. D., & Fox, J. (2015). Assessing the effects of gamification in the classroom: A longitudinal study on intrinsic motivation, social comparison, satisfaction, effort, and academic performance. *Computers & Education*, *80*, 152-161.

Harks, B., Rakoczy, K., Hattie, J., Besser, M., & Klieme, E. (2014). The effects of feedback on achievement, interest and self-evaluation: the role of feedback"s perceived usefulness. *Educational Psychology*, *34(3)*, 269-290.

Harley, B., & Swain, M. (1984). The interlanguage of immersion students and its implications for second language teaching. In A. Davies, C. Criper, & A. Howatt (Eds.), *Interlanguage* (pp. 291-311). Edinburgh: Edinburgh University Press.

Harrison, V., Kemp, R., Brace, N., & Snelgar, R. (2020). *SPSS for Psychologists*. Red Globe Press.

Hayes, J. R., & Flower, L. (1980). Identifying the organization of writing processes. In L. W. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 3-30). Hillsdale, NJ: Erlbaum.

Hayes, J. R. (2012). Modeling and remodeling writing. *Written Communication*, *29*, 369-388.

Herrington, A., & Moran, C. (2001). What happens when machines read our students" writing. *College English*, *63(4)*, 480-499.

Herrington, A., & Stanley, S. (2012). Criterion: Promoting the standard. In A. B. Inoue & M. Poe (Eds.), *Race and writing assessment* (pp. 47-61). New York, NY: Peter Lang.

Hirvela, A., Hyland, K., & Manchón, R. M. (2016). Dimensions in L2 writing theory and research: Learning to write and writing to learn. In R. M. Manchón & P. K. Matsuda (Eds.), *Handbook of second and foreign language writing* (pp. 45-63). De Gruyter Berlin, Germany.

Hoang, G. T. L., & Kunnan, A. J. (2016). Automated essay evaluation for English language learners: A case study of MY Access. *Language Assessment Quarterly*, *13(4)*, 359-376.

Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Shum, S. B., Santos, O. C., Rodrigo, M. T., Cukurova, M., & Bittencourt, I. I. (2021). Ethics of AI in education: towards a community-wide framework. *International Journal of Artificial Intelligence in Education*, 1-23.

Hoon, T. B. (2006). Online automated essay assessment: Potentials for writing development. *Making a difference with Web technologies. Proceedings of AusWeb*, 230-234.

Huang, S. J. (2014). Automated versus human scoring: A case study in an EFL context. *Electronic Journal of Foreign Language Teaching*, *11*, 149-164.

Hyland, F., & Hyland, K. (2001). Sugaring the pill: Praise and criticism in written feedback. *Journal of second language writing*,

Hyland, F. (1998). The impact of teacher written feedback on individual writers. *Journal of Second Language Writing*, *7*, 255-286.

Hyland, F. (2010). Future directions in feedback on second language writing: Overview and research agenda. *IJES, International Journal of English Studies*, *10*, 171-182.

Hyland, K. (2013). Student perceptions of hidden messages in teacher written feedback.

Studies in Educational Evaluation, *39*, 180-187.

Hyland, K., & Hyland, F. (2006). Feedback on second language students" writing. *Language teaching*,

Ishikawa, S. (1995). Objective measurement of low-proficiency EFL narrative writing. *Journal of Second Language Writing*, 4*(1)*, 51-69.

Iskander, M., Kapila, V., & Karim, M. A. (2010). *Technological developments in education and automation*. Springer Science & Business Media.

Jang, H., Reeve, J., & Deci, E. L. (2010). Engaging students in learning activities: It is not autonomy support or structure but autonomy support and structure. *Journal of educational psychology*, *102(3)*, 588.

Jason, T. (2020). Teachers, AI grammar checkers, and the newest literacies: Emending writing pedagogy and assessment. *Digital Culture & Education*, *12(1)*, 26-51.

Jeon, E. H., & Kaya, T. (2006). Effects of L2 instruction on interlanguage pragmatic development. *Synthesizing research on language learning and teaching*, 165-211.

Jiang, L., & Yu, S. (2020). Appropriating automated feedback in L2 writing: Experiences of Chinese EFL student writers. *Computer Assisted Language Learning*, 1-25.

Jin, W. (2001). A quantitative study of cohesion in Chinese graduate students" writing: variations across genres and proficiency levels.

Johansson, V. (2009). Lexical diversity and lexical density in speech and writing: a developmental perspective. *Lund Working Papers in Linguistics*, *53*, 61-79.

Johnson, R. L., Penny, J., Gordon, B., Shumate, S. R., & Fisher, S. P. (2005). Resolving score differences in the rating of writing samples: Does discussion improve the accuracy of scores. *Language Assessment Quarterly: An International Journal*, *2*, 117-146.

Kakkonen, T., Myller, N., & Sutinen, E. (2004). Semi-Automatic evaluation features in computer-assisted essay assessment. *Cate*, 456-461.

Kang, E., & Han, Z. (2015). The efficacy of written corrective feedback in improving L2 written accuracy: A meta-analysis. *Modern Language Journal*, *99*, 1-18.

Kasper, G., & Rose, K. R. (2002). Pragmatic development in a second language.

Language Learning: A Journal of Research in Language Studies, *52*, 1.

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

Kepner, C. G. (1991). An experiment in the relationship of types of written feedback to the development of second-language writing skills. *The modern language journal*,

Kim, H. R., & Bowles, M. (2019). How Deeply Do Second Language Learners Process Written Corrective Feedback? Insights Gained From Think-Alouds. *Tesol Quarterly*, *53(4)*, 913-938.

Kirsner, K. (1994). Implicit processes in second language learning. Implicit and explicit learning of languages. In N. Ellis (Ed.), *Implicit and explicit learning of languages* (pp. 283-312). San Diego, CA: Academic Press.

Krashen, S. D. (1984). *Writing, research, theory, and applications*. Pergamon.

Kyle, K., & Crossley, S. A. (2018). Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *The Modern Language Journal*, 102*(2)*, 333-349.

Kyle, K. (2016). *Measuring syntactic development in L2 writing : Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication.*

Kyle, K., & Crossley, S. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing*, *34*, 12-24.

Kyle, K., & Crossley, S. (2017). Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing*, *34*, 513-535.

Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): version 2.0. *Behaviour research methods*, *50(3)*, 1030-1046.

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, *49*, 757-786.

Kyle, K., Crossley, S. A., & McNamara, D. S. (2016). Construct validity in TOEFL iBT speaking tasks: Insights from natural language processing. *Language Testing*, *33*,

319-340.

Lalande, J. F. (1982). Reducing composition errors: An experiment. *The Modern Language Journal*,

Lam, R. (2016). Assessment as learning: examining a cycle of teaching, learning, and assessment of writing in the portfolio-based classroom. *Studies in Higher Education*, *41*, 1900-1917.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied linguistics*, *16(3)*, 307-322.

Lavolette, E., Polio, C., & Kahng, J. (2015). The accuracy of computer-assisted feedback and students" responses to it. *Language, Learning & Technology*, *19(2)*, 50-68.

Leach, L. (2012). Optional self-assessment: some tensions and dilemmas. *Assessment & Evaluation in Higher Education*, *37(2)*, 137-147.

(2014). *The oxford handbook of qualitative research*. Oxford University Press.

Lee, I. (1997). ESL learners" performance in error correction in writing: Some implications for teaching. *System*, *25(4)*, 465-477.

Lee, I. (2004). Error correction in L2 secondary writing classrooms: The case of Hong Kong. *Journal of Second Language Writing*, *13(4)*, 285-312.

Lee, I. (2005). Error correction in the L2 writing classroom: What do students think. *TESL Canada Journal*, *22(2)*, 1-16.

Lee, I. (2008a). Student reactions to teacher feedback in two Hong Kong secondary classrooms. *Journal of Second Language Writing*, *17(3)*, 144-164.

Lee, I. (2009). Ten mismatches between teachers" beliefs and written feedback practice. *ELT journal*, *63(1)*, 13-22.

Lee, I. (2011). Working smarter, not working harder: Revisiting teacher feedback in the L2 writing classroom. *Canadian modern language review*,

Lee, I. (2014a). Feedback in writing: Issues and challenges. *Assessing Writing*,

Lee, I. (2014b). Revisiting teacher feedback in EFL writing from sociocultural perspectives. *TESOL Quarterly*, *48*, 201-213.

Lee, I. (2016). Teacher education on feedback in EFL writing: Issues, challenges, and future directions. *Tesol Quarterly*,

Lee, I. (2017). *Classroom writing assessment and feedback in L2 school contexts*. Springer.

Lee, I. (2008b). Understanding teachers" written feedback practices in Hong Kong secondary classrooms. *Journal of second language writing*,

Lee, I., Yu, S., & Liu, Y. (2018). Hong Kong secondary students" motivation in EFL writing: A survey study. *Tesol Quarterly*, *52(1)*, 176-187.

Li, J., Link, S., & Hegelheimer, V. (2015). Rethinking the role of automated writing evaluation (AWE) feedback in ESL writing instruction. *Journal of Second Language Writing*, *27*, 1-18.

Li, Z., Feng, H. H., & Saricaoglu, A. (2017). The short-term and long-term effects of awe feedback on ESL students" development of grammatical accuracy. *Calico Journal*, 34*(3)*, 355-375.

Li, Z., Link, S., Ma, H., Yang, H., & Hegelheimer, V. (2014). The role of automated writing evaluation holistic scores in the ESL classroom. *System*, *44*, 66-78.

Lightbown, P. M., & Spada, N. (1994). An innovative program for primary ESL students in Quebec. *TESOL quarterly*, *28(3)*, 563-579.

Lim, F. V., & Phua, J. (2019). Teaching writing with language feedback technology. *Computers and Composition*, *54*, 1-13.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Newbury Park: Sage.

Ling, G., Elliot, N., Burstein, J. C., McCaffrey, D. F., MacArthur, C. A., & Holtzman, S. (2021). Writing motivation: A validation study of self-judgment and performance. *Assessing Writing*, *48*, 100509.

Link, S., Dursun, A., Karakaya, K., & Hegelheimer, V. (2014). Towards better ESL practices for implementing automated writing evaluation. *Calico Journal*, *31(3)*, 323-344.

Link, S. M. (2015). *Development and validation of an automated essay scoring engine to assess students" development across program levels.*

Liu, M., & Braine, G. (2005). Cohesive features in argumentative writing produced by Chinese undergraduates. *System*, 33*(4)*, 623-636.

Liu, S., & Kunnan, A. J. (2016). Investigating the application of automated writing

evaluation to Chinese undergraduate English majors: A case study of WriteToLearn. *calico journal*, *33(1)*, 71-91.

Long, M. (1996). The role of the linguistic environment in second language acquisition. *Handbook of second language acquisition*,

Long, M. H. (1981). Input, interaction, and second-language acquisition. *Annals of the New York academy of sciences*, *379(1)*, 259-278.

Lotherington, H., & Jenson, J. (2011). Teaching multimodal and digital literacy in L2 settings: New literacies, new basics, new pedagogies. *Annual Review of Applied Linguistics*, *31*, 226-246.

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers" language development. *TESOL quarterly*, *45(1)*, 36-62.

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners" oral narratives. *The Modern Language Journal*, *96(2)*, 190-208.

Mackey, A., Abbuhl, R., & Gass, S. M. (2012). Interactionist approach. In S. M. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 7-23). Routledge New York.

Maclellan, E. (2001). Assessment for learning: The differing perceptions of tutors and students. *Assessment & Evaluation in Higher Education*, *26(4)*, 307-318.

Manchón, R. M. (2009). *Writing in foreign language contexts: Learning, teaching, and research*. Multilingual Matters.

Mao, S. S., & Crosthwaite, P. (2019). Investigating written corrective feedback:(Mis) alignment of teachers" beliefs and practice. *Journal of Second Language Writing*, *45*, 46-60.

Mao, Z., & Lee, I. (2020). Feedback scope in written corrective feedback: Analysis of empirical research in L2 contexts. *Assessing Writing*, *45*, 100469.

Marshall, C., & Rossman, G. B. (2014). *Designing qualitative research*. Sage publications.

Mazgutova, D., & Kormos, J. (2015). Syntactic and lexical development in an intensive English for Academic Purposes programme. *Journal of Second Language Writing*,

29, 3-15.

McCutchen, D., & Perfetti, C. A. (1982). Coherence and connectedness in the development of discourse production. *ERIC*,

McDonald, S. A., & Shillcock, R. C. (2001). Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech*, *44(3)*, 295-322.

McNamara, D. S., Crossley, S. A., & Roscoe, R. (2013). Natural language processing in an intelligent writing strategy tutoring system. *Behaviour Research Methods*, *45*, 499-515.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (2010). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes*, *47*, 292-330.

Meara, P., & Bell, H. (2001). P_Lex: A simple and effective way of describing the lexical characteristics of short L2 texts. *Prospect*, *16(3)*, 5-19.

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

Morgan, H. (2016). Relying on high-stakes standardized tests to evaluate schools and teachers: A bad idea. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, *89(2)*, 67-72.

Moseley, M. H. (2006). Creating recursive writers in middle school: the effect of a writing program on student revision practices. *Unpublished Doctoral dissertation. Capella University, USA*,

Myles, J. (2002). Second language writing and research: The writing process and error analysis in student texts. *Tesl-Ej*, *6(2)*, 1-20.

Nassaji, H., & Swain, M. (2000). A Vygotskian perspective on corrective feedback in L2: The effect of random versus negotiated help on the learning of English articles. *Language awareness*, *9(1)*, 34-51.

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied linguistics*, *30(4)*, 555-578.

Nova, M., & Lukmana, I. (2018, July). *The detected and undetected errors in automated writing evaluation program"s result.* In English Language and Literature International Conference (ELLiC) Proceedings, 2 (pp. 120-126).

Nurmukhamedov, U., & Kim, S. H. (2009). "Would you perhaps consider…": hedged comments in ESL writing. *ELT journal*, *64(3)*, 272-282.

Nystrom, N. (1983). Teacher-student interaction in bilingual classrooms: Four approaches to error feedback. In H. Seliger & M. H. Long (Eds.), *Classroom oriented research in second language acquisition* (pp. 169-189). Rowley, MA: Newbury House.

O"Neill, R., & Russell, A. (2019). Grammarly: Help or hindrance? Academic learning advisors" perceptions of an online grammar checker. *Journal of Academic Language and Learning*, *13(1)*, A88-A107.

O"reilly, T., & McNamara, D. S. (2007). Reversing the reverse cohesion effect: Good texts can be better for strategic, high-knowledge readers. *Discourse processes*, *43(2)*, 121-152.

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, *24*, 492-518.

Ortega, L. (2012). Epilogue: Exploring L2 Writing-SLA interfaces. *Journal of Second Language Writing*, *21*, 404-415.

Oxford, R. L. (1990). *Language Learning Strategies*. New York: Heinle & Heinle.

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

Park, J. H., & Yang, I. Y. (2020). Utilizing an AI-Based grammar checker in an EFL writing classroom. 응용언어학, *36(1)*, 97-120.

Patton, M. Q. (1987). *How to use qualitative methods in evaluation* ((4)). Sage.

Perelman, L. (2012). Construct validity, length, score, and time in holistically graded writing assessments: The case against automated essay scoring (AES). In C.

Bazerman, C. Dean, J. Early, K. Lunsford, S. Null, P. Rogers, & A. Stansell (Eds.), *International advances in writing research: Cultures, places, measures* (pp. 121-131). Fort Collins, Colorado/Anderson, SC: WAC Clearinghouse/Parlor Press.

Perl, S. (1979). The composing processes of unskilled college writers. *Research in the Teaching of English*, *13(4)*, 317-336.

Petchprasert, A. (2021). Utilizing an automated tool analysis to evaluate EFL students" writing performances. *Asian-Pacific Journal of Second and Foreign Language Education*, *6(1)*, 1-16.

Peterson, M. (2010). Computerized games and simulations in computer-assisted language learning: A meta-analysis of research. *Simulation & Gaming*,

Polio, C. (2012). The relevance of second language acquisition theory to the written error correction debate. *Journal of Second Language Writing*, *21*, 375-389.

Polio, C. (2017). Second language writing development: A research agenda. *Language Teaching*, *50*, 261-275.

Prensky, M. (2001). Digital natives, digital immigrants. *On the horizon*, *9(5)*, 1-9.

Qin, J. (2014). Use of formulaic bundles by non-native English graduate writers and published authors in applied linguistics. *System*, *42*, 220-231.

Qin, W., & Uccelli, P. (2016). Same language, different functions: A cross-genre analysis of Chinese EFL learners" writing performance. *Journal of Second Language Writing*, *33*, 3-17.

Quinton, S., & Smallbone, T. (2010). Feeding forward: using feedback to promote student reflection and learning–a teaching model. *Innovations in Education and Teaching International*, *47(1)*, 125-135.

Ramineni, C., & Williamson, D. (2013). Automated essay scoring: Psychometric guidelines and practices. *Assessing Writing*, *18*, 25-39.

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

Ranalli, J. (2018). Automated written corrective feedback: how well can students make use of it. *Computer Assisted Language Learning*, *31(7)*, 653-674.

Rassaei, E. (2014). Scaffolded feedback, recasts, and L2 development: A sociocultural perspective. *The Modern Language Journal*, *98(1)*, 417-431.

Read, J. (2000). *Assessing Vocabulary*. Cambridge University Press.

Reid, J. (1994). Responding to ESL students" texts: The myths of appropriation. *Tesol Quarterly*, *28(2)*, 273-292.

Robb, T., Ross, S., & Shortreed, I. (1986). Salience of feedback on error and its effect on EFL writing quality. *TESOL quarterly*, *20(1)*, 83-96.

Roed, J. (2003). Language learner behaviour in a virtual environment. *Computer assisted language learning*, *16(2-3)*, 155-172.

Sadeghi, K., & Rahmati, T. (2017). Integrating assessment as, for, and of learning in a large-scale exam preparation course. *Assessing Writing*, *34*, 50-61.

Schiftner, B. (2013). Analysing coherence in upper-intermediate learner writing. In A. Diaz-Negrillo, N. Ballier, & P. Thompson (Eds.), Automatic treatment and analysis of
learner corpus data (pp. 265-228). Amsterdam: John Benjamins Publishing.

Schmidt, R. (1995). Consciousness and foreign language learning: A tutorial on the role of attention and awareness in learning. *Attention and awareness in foreign language learning*, *9*, 1-63.

Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3-32). Cambridge: Cambridge University Press.

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

Scott, V. M. (1996). *Rethinking foreign language writing*. Heinle & Heinle Publishers.

Semke, H. D. (1984). Effects of the red pen. *Foreign language annals*,

Seppälä, J., Mitsuishi, T., Ohkawa, Y., Zhao, X., & Nieminen, M. (2020). *Study on UX design in enhancing student motivations in mobile language learning.* In 2020 IEEE International Conference on Teaching, Assessment, and Learning for

Engineering (TALE) (pp. 948-951).

Shao, X. (2015). On written corrective feedback in L2 writing. *English Language Teaching*, *8*, 155-168.

Shaw, S. D., & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing* (26). Cambridge University Press.

Sheen, Y. (2008). Recasts, language anxiety, modified output, and L2 learning. *Language Learning*, *58(4)*, 835-874.

Sheen, Y., Wright, D., & Moldawa, A. (2009). Differential effects of focused and unfocused written correction on the accurate use of grammatical forms by adult ESL learners. *System*, *37(4)*, 556-569.

Sheen, Y. (2007). The effect of focused written corrective feedback and language aptitude on ESL learners" acquisition of articles. *TESOL Quarterly*, *41*, 255-283.

Sheen, Y. (2010). Introduction: The role of oral and written corrective feedback in SLA. *Studies in Second Language Acquisition*, *32*, 169-179.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational researcher*, *29(7)*, 4-14.

Sheppard, K. (1992). Two feedback types: Do they make a difference? *RELC journal*, *23(1)*, 103-110.

Shermis, M. D., & Burstein, J. (Eds.). (2003). *Automated essay scoring: A cross-disciplinary perspective*. New Jersey: Mahwah.

Shermis, M. D., Garvan, C. W., & Diao, Y. (2008a). The impact of automated essay scoring on writing outcomes. *Online Submission*, 45.

Shermis, M. D., & Hamner, B. (2012, April). *Contrasting state-of-the-art automated scoring of essays: Analysis.* In Annual national council on measurement in education meeting.

Shermis, M. D., Shneyderman, A., & Attali, Y. (2008b). How important is content in the ratings of essay assessments? *Assessment in Education: Principles, Policy & Practice*, *15*, 91-105.

Shintani, N. (2016). The effects of computer-mediated synchronous and asynchronous direct corrective feedback on writing: a case study. *Computer Assisted Language*

Learning, *29*, 517-538.

Sommers, N. (1980). Revision strategies of student writers and experienced adult writers. *College composition and communication*, *31(4)*, 378-388.

Spandel, V. (2005). *Creating writers: Through 6-trait writing assessment and instruction*. Pearson Allyn and Bacon.

Stallard, C. K. (1974). An analysis of the writing behaviour of good student writers. *Research in the Teaching of English*, *8(2)*, 206-218.

Staples, S., Egbert, J., Biber, D., & McClair, A. (2013). Formulaic sequences and EAP writing development: Lexical bundles in the TOEFL iBT writing section. *Journal of English for academic purposes*, *12(3)*, 214-225.

Staples, S., Egbert, J., Biber, D., & Gray, B. (2016). Academic writing development at the university level: Phrasal and clausal complexity across level of study, discipline, and genre. *Written Communication*, *33(2)*, 149-183.

Stefanou, C., & Révész, A. (2015). Direct written corrective feedback, learner differences, and the acquisition of second language article use for generic and specific plural reference. *Modern Language Journal*, *99*, 263-282.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, *9(4)*, 1-19.

Sternberg, R. J. (1992). CAT: A program of comprehensive abilities testing. In B. R. Gifford & M. C. O"Connor (Eds.), *Changing Assessments* (pp. 213-274). Boston:: Kluwer Academic Publishers.

Stevenson, M. (2016). A critical interpretative synthesis: The integration of Automated Writing Evaluation into classroom writing instruction. *Computers and Composition*, *42*, 1-16.

Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. *Assessing Writing*, *19*, 51-65.

Stevenson, M., & Phakiti, A. (2019). Automated feedback and second language writing. *Feedback in second language writing: Contexts and issues*, 125-142.

Storch, N. (2009). The impact of studying in a second language (L2) medium university

on the development of L2 writing. *Journal of Second Language Writing*, *18*, 103-118.

Storch, N. (2010). Critical feedback on written corrective feedback research. *International Journal of English Studies*, *10*, 29-46.

Storch, N. (2018). Written corrective feedback from sociocultural theoretical perspectives: A research agenda. *Language Teaching*, *51*, 262-277.

Storch, N., & Tapper, J. (2009). The impact of an EAP course on postgraduate writing. *Journal of English for Academic Purposes*, *8*, 207-223.

Storch, N., & Wigglesworth, G. (2010). Learners" processing, uptake, and retention of corrective feedback on writing. *Studies in Second Language Acquisition*, *32*, 303-334.

Sugita, Y. (2006). The impact of teachers" comment types on students" revision. *ELT journal*, *60(1)*, 34-41.

Sung, Y.-T., Chang, K.-E., & Liu, T.-C. (2016). The effects of integrating mobile devices with teaching and learning on students" learning performance: A meta-analysis and research synthesis. *Computers & Education*, *94*, 252-275.

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. M. Gass & C. Madden (Eds.), *Input in second language acquisition* (pp. 235-253). Rowley, MA: Newbury House.

Swain, M. (1993). The output hypothesis: Just speaking and writing aren"t enough. *Canadian modern language review*, *50(1)*, 158-164.

Swain, M. (1995). *Three functions of output in second language learning. principles and practice in applied linguistics: studies in honor of H. G. Widdowson.* Oxford: Oxford University Press.

Swain, M. (2000). The output hypothesis and beyond: Mediating acquisition through collaborative dialogue. In J. Lantolf (Ed.), *Sociocultural theory and second language learning* (pp. 97-114). Oxford: Oxford University Press.

Swain, M., & Lapkin, S. (2002). Talking it through: Two French immersion learners" response to reformulation. *International Journal of Educational Research*,

37(3-4), 285-304.

Teddlie, C., & Tashakkori, A. (2010). Overview of contemporary issues in mixed methods research. In *SAGE Handbook of Mixed Methods in Social & Behavioural Research* (pp. 1-42). Sage.

Thumlert, K., de Castell, S., & Jenson, J. (2015). Short cuts and extended techniques: Rethinking relations between technology and educational theory. *Educational Philosophy and Theory*, *47*, 786-803.

Thumlert, K., Owston, R., & Malhotra, T. (2018). Transforming school culture through inquiry-driven learning and iPads. *Journal of Professional Capital and Community*,

Tian, L., & Zhou, Y. (2020). Learner engagement with automated feedback, peer feedback and teacher feedback in an online EFL writing context. *System*, *91*, 102247.

Tocalli-Beller, A., & Swain, M. (2005). Reformulation: The cognitive conflict and L2 learning it generates. *International Journal of Applied Linguistics*, *15(1)*, 5-28.

Truscott, J. (1996). The case against grammar correction in L2 writing classes. *Language learning*, *46(2)*, 327-369.

Truscott, J., & Hsu, A. Y. (2008). Error correction, revision, and learning. *Journal of second language writing*, *17(4)*, 292-305.

Truscott, J. (2007). The effect of error correction on learners" ability to write accurately. *Journal of Second Language Writing*, *16*, 255-272.

Ursachi, G., Horodnic, I. A., & Zait, A. (2015). How reliable are measurement scales? External factors with indirect influence on reliability estimators. *Procedia Economics and Finance*, *20*, 679-686.

Van Beuningen, C. (2010). Corrective feedback in L2 writing: Theoretical perspectives, empirical insights, and future directions. *International Journal of English Studies*, *10*, 1-27,184.

Van Beuningen, C. G. (2011). *The effectiveness of comprehensive corrective feedback in second language writing*. Oisterwijk: Uitgeverij BOXPress.

Van Beuningen, C. G., De Jong, N. H., & Kuiken, F. (2012). Evidence on the

effectiveness of comprehensive error correction in second language writing. *Language learning*, *62(1)*, 1-41.

Verspoor, M., & Lowie, W. (2003). Making sense of polysemous words. *Language learning*, *53(3)*, 547-586.

Vidakovic, I., & Barker, F. (2010). Use of words and multi-word units in Skills for Life Writing examinations. *Cambridge ESOL: Research Notes*, *41*, 7-14.

Vygotsky, L. S. (1980). *Mind in society: The development of higher psychological processes*. Harvard university press.

Vygotsky, L. S. (1987). *The collected works of LS Vygotsky: Vol. 1, Problems of general psychology (RW Rieber & AS Carton, Eds., N. Minick, trans.)*. New York: Plenum Press.

Wang, J., & Brown, M. S. (2008). Automated essay scoring versus human scoring: A correlational study. *Contemporary Issues in Technology and Teacher Education*, *8*, 310-325.

Wang, M.-J., & Goodman, D. (2012). Automated writing evaluation: Students" perceptions and emotional involvement. *English Teaching & Learning*, *36*, 1-37.

Wang, P. (2015). Effects of an automated writing evaluation program: Student experiences and perceptions. *Electronic Journal of Foreign Language Teaching*,

Wang, P.-l. (2013). Can automated writing evaluation programs help students improve their english writing. *International Journal of Applied Linguistics and English Literature*, *2(1)*, 6-12.

Wang, T., & Jiang, L. (2015). Studies on written corrective feedback: Theoretical perspectives, empirical evidence, and future directions. *English Language Teaching*, *8(1)*, 110-120.

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

Ware, P. (2011). Computer-generated feedback on student writing. *TESOL Quarterly*, *45*, 769-774.

Ware, P. D., & Warschauer, M. (2006). Electronic feedback and second language writing.

In K. H. Hyland, Fiona (Ed.), *Feedback in second language writing: Contexts and issues* (pp. 105-122). Cambridge University Press New York.

Warschauer, M. (2010). Invited commentary: New tools for teaching writing. *Language Learning & Technology*, *14*, 3-8.

Warschauer, M., & Grimes, D. (2008). Automated writing assessment in the classroom. *Pedagogies: An International Journal*, *3(1)*, 22-36.

Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language teaching research*, *10(2)*, 157-180.

Weigle, S. C. (2010). Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability. *Language Testing*, *27*, 335-353.

Weigle, S. C. (2013a). English as a second language writing and automated essay evaluation. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 36-54). Routledge.

Weigle, S. C. (2013b). English language learners and automated scoring of essays: Critical considerations. *Assessing Writing*, *18*, 85-99.

Weir, C. (2005). *Language testing and validation: an evidence-based approach*. Palgrave Macmillan.

Wigglesworth, G., & Storch, N. (2012). What role for collaboration in writing and writing feedback? *Journal of Second Language Writing*, *21*, 364-374.

Williams, J. (2012). The potential role(s) of writing in second language development. *Journal of Second Language Writing*, *21*, 321-331.

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*, 691-718.

Wolfe-Quintero, K., Inagaki, S., Kim, H.-Y., Kim, H.-Y., & Inagaki, S. (1998). *Second language development in writing: Measures of fluency, accuracy, & complexity*. University of Hawai"i Press Honolulu.

Wolter, B. (2001). Comparing the L1 and L2 mental lexicon: A depth of individual word knowledge model. *Studies in second language acquisition*, 41-69.

Woodworth, J., & Barkaoui, K. (2020). Perspectives on using automated

writing evaluation systems to provide written corrective feedback in the ESL classroom. *TESL Canada Journal*, *37(2)*, 234-247.

Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing*, *27(3)*, 291-300.

Xu, C. (2009). Overgeneralization from a narrow focus: A response to Ellis et al. (2008) and Bitchener (2008). *Journal of Second Language Writing*, *18*, 270-275.

Yang, N. D. (2004). *Using My Access in EFL writing.* In The proceedings of 2004 international conference and workshop on TESL & applied linguistics, (pp. 550-564).

Yang, W., Lu, X., & Weigle, S. C. (2015). Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. *Journal of Second Language Writing*, *28*, 53-67.

Yannakoudakis, H., Andersen, Ø. E., Geranpayeh, A., Briscoe, T., & Nicholls, D. (2018). Developing an automated writing placement system for ESL learners. *Applied Measurement in Education*, *31*, 251-267.

Yeh, S.-W., & Lo, J.-J. (2009). Using online annotations to support error correction and corrective feedback. *Computers & Education*, *52(4)*, 882-892.

Yu, Y. T., & Yeh, Y. L. (2003). *Computerized feedback and bilingual concordancer for EFL college students" writing.* In Proceedings of the 2003 International Conference on English Teaching and Learning in the Republic of China (pp. 35-48).

Zamel, V. (1985). Responding to student writing. *TESOL Quarterly*, *19(1)*, 79-101.

Zamin, A. A. M. (2021). The Use of Automated Writing Evaluation (AWE) in Developing Language Proficiency: A Study from the Learners" Perspective. *Malaysian Journal of ELT Research*, *17(2)*, 87-102.

Zhang, M. (2013). Contrasting automated and human scoring of essays. *R & D Connections*, *21(2)*, 1-11.

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

Zhang, Z. (2020). Engaging with automated writing evaluation (AWE) feedback on L2

writing: Student perceptions and revisions. *Assessing Writing*, *43*, 1-14.

Zhang, Z. V., & Hyland, K. (2018). Student engagement with teacher and automated feedback on L2 writing. *Assessing Writing*, *36*, 90-102.

Zhao, H. (2010). Investigating learners" use and understanding of peer and teacher feedback on writing: A comparative study in a Chinese English writing classroom. *Assessing writing*, *15(1)*, 3-17.

Zhu, M., Liu, O. L., & Lee, H.-S. (2020). The effect of automated feedback on revision behaviour and learning gains in formative assessment of scientific argument writing. *Computers & Education*, *143*, 1-15.

Appendices

Appendix A: Description of the ETS Criterion Score Guide

ETS Criterion Score Guide from https://Criterion.ets.org/Content/topics/topics-toefl.htm

Score = 6
A typical essay at this level:
- effectively addresses the writing task
- is well organized and well developed
- uses clearly appropriate details to support a thesis or illustrate ideas
- displays consistent facility in the use of language
- demonstrates syntactic variety and appropriate word choice, though it may have occasional errors

Score = 5
A typical essay at this level:
- may address some parts of the task more effectively than others
- is generally well-organized and well-developed
- uses details to support a thesis or illustrate idea
- displays facility in the use of language
- demonstrates some syntactic variety and range of vocabulary, though it will probably have occasional errors

Score = 4
A typical essay at this level:
- addresses the writing topic adequately but may slight parts of the task
- is adequately organized and developed
- uses some details to support a thesis or illustrate an idea
- demonstrates adequate but possibly inconsistent facility with syntax and usage
- may contain some errors that occasionally obscure meaning

Score = 3
A typical essay at this level may reveal one or more or the following weaknesses:
- inadequate organization or development
- inappropriate or insufficient details to support or illustrate generalizations
- a noticeably inappropriate choice of words or word forms
- an accumulation of errors in sentence structure and/or usage

Score = 2
A typical essay at this level is flawed by one or more of the following weaknesses:
- serious disorganization or underdevelopment
- little or no detail, or irrelevant specifics
- serious and frequent errors in sentence structure and usage
- serious problems with focus

Score = 1
A typical essay at this level:

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

Appendix B: Student Background and Perception Questionnaire

Student Background and Perception Questionnaire

Thank you for participating in this project!

Please fill out some information before we begin. As you have been informed, your responses will be confidential (kept secret) and will only be used for research purposes.

ID (Given to each student) _____
Which country are you from? _____
What is your first language? _____
Gender _____
Age _____
What is your IELTS or TOEFL writing score? _____
How many years have you studied English? _____
How long have you been in Canada? _____
What is your intended major? _____

Please answer the questions about your writing practices in the past.
Definitely agree 4, Mostly Agree 3, Mostly Disagree 2, Definitely Disagree 1

1	I think doing more writing is important to improve my writing.
2	I pay attention to the score when my writing is returned.
3	I pay attention to the feedback when my writing is returned.
4	I think the feedback I received from my teachers was timely.
5	I try to avoid similar problems in future writing when I receive feedback.
6	I revise my essays before submission.
7	I think revising my essays is an important part of the writing process.
8	I like revising my essays.
9	I find teacher feedback helpful when revising my essays.
10	I find peer feedback helpful in revising my essays.
11	I have previous experience with computer feedback systems (e.g. Grammarly, Microsoft word grammar, spelling checked, turnitin.com, etc…)
12	If yes to 11, I find computer feedback helpful in revising my essays.

Appendix C: Writing Process Questionnaire

Writing Process Questionnaire Pre-treatment
Adapted from Chan et al. (2017)

ID (Given to each student) _____

Definitely agree 4, Mostly Agree 3, Mostly Disagree 2, Definitely Disagree 1

While weading an essay prompt….	
1	I usually read the whole prompt carefully
2	I usually think about how well I understood the task requirements.
3	I usually think about what I know about the topic.
4	I usually think about what I know about the genre.
5	I usually think about the purpose of the task
6	I usually think about what I might need to write to make my essay relevant and adequate to the task.
7	I usually think about the intended reader of my essay and their expectations.
8	I usually think about or jot down ideas which are relevant to the task/topic.
9	I usually prioritize my ideas based on the task requirements.
10	I usually link my ideas based on the task requirements.
11	I usually work out how my ideas relate to each other, e.g. main ideas or examples.
12	I usually think about the structure of my essay.
13	I usually remove some ideas I planned to write because they did not fit the structure of my essay.
14	I usually re-read the prompt/task instructions.

Please answer the questions about your writing practices in general for writing academic essays.

About your writing Practices in general for writing academic essays

While writing the first draft….	
15	I usually think about the appropriate words to express my ideas.
16	I usually think about the correct sentence structures to express my ideas.
17	I usually think about the correct grammar to express my ideas
18	I usually think about how to connect my ideas smoothly in the whole essay
19	I usually think about how to make my ideas persuasive to the reader.
20	I usually organize my sentences and paragraphs in a logical order.
21	I usually brain-storm main and supporting ideas.
22	I usually re-read the task instructions/prompts.
23	I usually change my writing plan (e.g. structure and content)
24	I usually check that the content is relevant and revise accordingly.
25	I usually check that my essay is well-organized and revise accordingly.

26	I usually check that my essay is coherent and revise accordingly.
27	I usually check that I include my own viewpoint on the topic and revise accordingly.
28	I usually check the possible effect of my essay on the intended reader and revise accordingly.
29	I usually check the accuracy and range of sentence structures and revise accordingly.
30	I usually check the grammar (e.g. part of speech and tenses) and revise accordingly.
31	I usually check the appropriateness and range of vocabulary and revise accordingly.

After writing the first draft ….	
32	I usually write multiple drafts.
If "definitely agree" or "mostly agree" to the previous question, move on to the following questions.	
33	I usually check that the content is relevant and revise accordingly.
34	I usually check that my essay is well organized and revise accordingly.
35	I usually check that my essay is coherent and revise accordingly.
36	I usually check that I include my own viewpoint on the topic and revise accordingly.
37	I usually check the possible effect of my essay on the intended reader and revise accordingly.
38	I usually check the accuracy and range of the sentence structures and revise accordingly.
39	I usually check the grammar (e.g. part of speech and tenses) and revise accordingly.
40	I usually check the appropriateness and range of vocabulary and revise accordingly.

Writing Process Questionnaire Post-treatment

ID (Given to each student) _____

Please answer the questions about your writing practices in the course after using computer feedback in the last writing task (in-class writing 3).

Definitely agree 4, Mostly Agree 3, Mostly Disagree 2, Definitely Disagree 1

While reading the essay prompt….	
1	I read the whole prompt carefully
2	I thought about how well I understood the task requirements.
3	I thought about what I know about the topic.
4	I thought about what I know about the genre.
5	I thought about the purpose of the task
6	I thought about what I might need to write to make my essay relevant and adequate to the task.
7	I thought about the intended reader of my essay and their expectations.
8	I thought about or jotted down ideas which are relevant to the task/topic.
9	I prioritized my ideas based on the task requirements.
10	I linked my ideas based on the task requirements.
11	I worked out how my ideas relate to each other (e.g. main ideas or examples).
12	I thought about the structure of my essay.
13	I removed some ideas I planned to write because they did not fit the structure of my essay.
14	I re-read the prompt/task instructions.

While writing the first draft….	
15	I thought about the appropriate words to express my ideas.
16	I thought about the correct sentence structures to express my ideas.
17	I thought about the correct grammar to express my ideas
18	I thought about how to connect my ideas smoothly in the whole essay
19	I thought about how to make my ideas persuasive to the reader.
20	I organized my sentences and paragraphs in a logical order.
21	I brain-stormed main and supporting ideas.
22	I re-read the task instructions/prompts.
23	I changed my writing plan (e.g. structure and content)
24	I checked that the content is relevant and revised accordingly.
25	I checked that my essay is well-organized and revised accordingly.
26	I checked that my essay is coherent and revised accordingly.
27	I checked that I include my own viewpoint on the topic and revised accordingly.
28	I checked the possible effect of my essay on the intended reader and revised accordingly.
29	I checked the accuracy and range of sentence structures and revised accordingly.
30	I checked the grammar (e.g. part of speech and tenses) and revised accordingly.

31	I checked the appropriateness and range of vocabulary and revised accordingly.

After receiving feedback on my essay….	
32	I checked that the content was relevant and revised accordingly.
33	I checked that my essay was well organized and revised accordingly.
34	I checked that my essay was coherent and revised accordingly.
35	I checked that I include my own viewpoint on the topic and revised accordingly.
36	I checked the possible effect of my essay on the intended reader and revised accordingly.
37	I checked the accuracy and range of the sentence structures and revised accordingly.
38	I checked the grammar (e.g. part of speech and tenses) and revised accordingly.
39	I checked the appropriateness and range of vocabulary and revised accordingly.

Appendix D: Perception of Criterion Questionnaire

Adapted from Dikli and Bleyle (2014)	**Strongly Disagree (1) (%)**	**Disagree (2) (%)**	**Neutral (3) (%)**	**Agree (4) (%)**	**Strongly agree (5) (%)**	**Mean**	**SD**
I found the scoring rubric of Criterion helpful to understand my ability.	17.65	17.65	58.82	5.88	0.00	2.53	0.87
I found the holistic score that Criterion provided adequate to understand my performance.	17.65	35.29	47.06	0.00	0.00	2.29	0.77
I found the trait score that Criterion provided adequate to understand my performance.	17.65	29.41	47.06	5.88	0.00	2.41	0.87
I found using Criterion feedback helpful in revising my essays.	0.00	0.00	11.76	76.47	11.76	4.00	0.50
I found the scoring speed of Criterion satisfying.	0.00	0.00	0.00	5.88	94.12	4.94	0.24
I found the Criterion feedback on Grammar helpful (e.g., run-on, agreement, pronoun errors)	0.00	11.76	11.76	52.94	23.53	3.88	0.93
I found the Criterion feedback on Usage helpful (e.g., article, word form, preposition error)	0.00	5.88	17.65	58.82	17.65	3.88	0.78
I found the Criterion feedback on Mechanics helpful (e.g., spelling, capitalization, punctuation)	0.00	0.00	5.88	52.94	41.18	4.35	0.61
I found the Criterion feedback on Style helpful (e.g., repetition, passive, sentence length)	0.00	29.41	47.06	17.65	5.88	2.76	0.83
I found the Criterion feedback on Organization & Development helpful (e.g., thesis, ideas, conclusion)	0.00	29.41	35.29	35.29	0.00	3.06	0.83
I think my English writing ability has improved after using Criterion® .	0.00	5.88	23.53	52.94	17.65	3.82	0.81
I think Criterion is user-friendly.	5.88	17.65	29.41	29.41	17.65	3.35	1.17
I think I will use Criterion again in the future if I have the chance.	0.00	0.00	17.65	47.06	35.29	4.18	0.73
Generally speaking, I am satisfied with Criterion®.	5.88	11.76	23.53	41.18	17.65	3.53	1.12

Appendix E: Teacher Analytic Rubric

	1	**2**	**3**	**4**	**DY 80+**
Topic / Task Response	Unintelligible and/or too short for assessment	Topic is not addressed due to misunderstanding of the prompt or memorization. A great deal of irrelevant material included.	Candidate has made a reasonable attempt to accomplish the task. Most points in the essay prompt are addressed. Some irrelevant material may be included. Little evidence of independent thought. Little analysis of any of the ideas presented.	Candidate has addressed the all aspects of the topic, albeit simply. At least some aspects of topic are thoroughly developed and others at least adequately. Candidate has communicated some personal interest in the topic and provided relevant examples to support the argument	Full realization of task. Candidate has addressed all points in the essay prompt. All aspects of the topic are thoroughly developed. Essay incorporates interesting, original ideas.
Organization / Coherence & Cohesion	Unintelligible and/or too short for assessment	Little evidence of organization. No linking between ideas.	Essay has no clear introduction, body or conclusion. Repetitive and mechanical use of connectives and cohesive devices. These are often faulty. Referencing is used but often faulty.	Essay has a clear introduction, body and conclusion. Usually clear coherence and organization including topic sentences and logical sequencing of ideas. Connectives and cohesive devices are used to good effect with minor inappropriacies. Referencing is used with minor inappropriacies.	Logical and well organized. Candidate presents points and evidence with clarity and conciseness and achieves a logical flow of ideas. Coherence and Cohesion are well preserved with skillful usage of connective and cohesive devices. Referencing is used skillfully.

Lexical Range & Accuracy	Unintelligible and/or too short for assessment	Lack of range in vocabulary makes performance inadequate. Inappropriate use of many lexical items. Meaning is greatly obscured.	Range of vocabulary is basic but usually adequate for the task. Little variety of expression. Some paraphrasing and repetition. Mistakes in word forms and spelling impedes communication but can be understood with effort.	Solid vocabulary resource. Adequate vocabulary despite some gaps in more specialized areas and in idiomatic usage. Minor mistakes in word forms and spelling rarely impedes communication.	Sophisticated, wide-ranging vocabulary used appropriately in all contexts.
Grammatical Range & Accuracy	Unintelligible and/or too short for assessment	Little control of basic structures. Frequent elementary errors. Meaning is greatly obscured.	Adequate grasp of basic structures. Candidate makes some attempt to use complex structures but is rarely successful. Grammatical accuracy impedes communication but can be understood with effort.	Generally accurate use of a range of structures. Some occasional minor lapses. Writer's meaning is clear despite some common errors. Grammatical accuracy rarely impedes communication.	Sophisticated syntax approaching native-speaker level. Minimal errors. Grammatical accuracy does not impede communication.

Appendix F: Rating Scale for Revision

Adapted from Ferris (1997)

0	No discernible change made by student in response to this comment
1	Minimal attempt by student to address the comment, effect generally negative or negligible
2	Substantive change(s) made by student in response to the comment, effect generally negative or negligible
3	Minimal attempt by student to address the comment, effect mixed
4	Substantive change(s) made by student in response to the comment, effect mixed
5	Minimal attempt by student to address the comment, effect generally positive
6	Substantive change(s) made by student in response to the comment, effect generally positive

Appendix G: Semi-Structured Questions for Student Focus-Group Interviews

Semi-Structured Questions for Student Focus-Group Interviews
Adapted from Li et al. (2015)

Opening Script:
The purpose of this interview is to learn about how effective hybrid feedback was for developing your writing. There are no right or wrong answers, or desirable or undesirable answers. I would like you to feel comfortable saying what you really think and how you really feel. If it"s okay with you, I will be recording the conversation since it is hard for me to write down everything while simultaneously carrying an attentive conversation with you.

Perception of AWE feedback

- How do you feel about the use of computer feedback in your writing?
- Was the following error analysis useful?
- grammar (e.g., subject verb agreement, ill-formed verb)
- usage (e.g., article, preposition, word choice)
- mechanics (e.g., spelling, punctuation, capitalization)
- style (e.g., repeated words, long/short sentences, passive sentences)

- What kind of strategies did you use to achieve the Criterion holistic score of 4?
- Do you think the Criterion scores and feedback were objective?
- Did having immediate feedback help you to be more motivated to revise?
- Is it easier now to find/identify errors by yourself after using Criterion?
- Can you identify your writing weaknesses from the feedback in Criterion?
- Did you proofread/revise or correct any of your mistakes by yourself before you submitted the paper to Criterion and get to see what was wrong with your paper? If "yes", Why? / If "no", why not?
- While you have been using Criterion in your class, what was the most impressive/ interesting points while using Criterion?
- Which part of writing do you think has improved as a result of using Criterion? (e.g., grammar/organization/wording/spelling)
- How did you feel about your writing before submitting it to your teacher?
- How did you feel about using Criterion in terms of self-efficacy, meta-cognition, confidence, and self-satisfaction?
- Did using AWE for feedback feel like a "game" in any way?
- Do you have any other experiences or suggestions you would like to share concerning Criterion feedback?

Perception of teacher feedback

- Do you see any differences between the teacher"s feedback on your paper and Criterion"s feedback? If "yes", please explain the differences between them.
- Do you find teacher feedback on content and organization useful?
- Do you find teacher feedback on grammar, usage, mechanics and style more useful than Criterion"s?
- Do you think teacher feedback and scores are more objective than Criterion scores and feedback?
- Did the delay in having feedback affect your motivation to revise?
- Do you think Criterion feedback can replace teacher feedback?
- Do you have any other experiences or suggestions you would like to share concerning teacher feedback?

Perception of hybrid feedback

- Do you think combining teacher and Criterion feedback is more effective than teacher feedback alone?
- Do you think combining teacher and Criterion feedback is more effective than Criterion feedback alone?
- Do you think the hybrid feedback enhanced your autonomy (independence) as a learner?
- Do you have any other experiences or suggestions you would like to share concerning combining teacher and Criterion feedback?
- Did the combination of feedback motivate you to revise your writing more?
- After your first hybrid feedback, did you focus more on any of the following:
- grammar (e.g., subject verb agreement, ill-formed verb)
- usage (e.g., article, preposition, word choice)
- mechanics (e.g., spelling, punctuation, capitalization)
- style (e.g., repeated words, long/short sentences, passive sentences)
- content (e.g., quality of argumentation, supporting ideas)
- organization (e.g., topic sentences, paragraphing, connecting ideas)
- After this experiment, do you see writing more as a process or as a product?
- Walk me through your writing process before the experiment and after the experiment.

Closing script:

Before we wrap things up and talk about next steps, are there any last comments you have regarding this interview?

In about a week later, I"ll send you a copy of the transcript of this interview for your review. If you would like to add any clarifications, please let me know within a week"s time.

Thank you for your participation, and please do not hesitate to contact me should you think of additional areas that we should include or if you have any questions.

Appendix H: Informed Consent Forms

Informed Consent Form – Experimental group

Date:

Study Name: Hybrid Feedback: The Efficacy of Combining Automated Writing Corrective Feedback and Teacher Feedback for Academic Writing Development in an English for Academic Purpose (EAP) Context

Researcher name:

Johanathan Woodworth - PhD program in Education at York University (Principle investigator)
Johanathan_Woodworth@edu.yorku.ca

Khaled Barkaoui (Supervisor)
KBarkaoui@edu.yorku.ca
(416) 736 2100 ext. 33209

Purpose of the Research:

The purpose of the research is to investigate the effectiveness of teacher and machine feedback compared to teacher feedback only on student"s improvement in writing. The research will specifically examine language, content and organization of the writing.

The research will be conducted by collecting questionnaire, interview data and collecting your writing samples for analysis. The findings will be reported in a PhD dissertation.

What You Will Be Asked to Do in the Research:

1. You will complete 2 questionnaires at the beginning of the class and 1 at the end of the semester.
   a. In the beginning of the semester
      i. Background and perception of feedback questionnaire (~10 minutes)
      ii. Writing Process Questionnaire(~30 minutes)
   b. At the end of the session
      i. Writing Process Questionnaire(~30 minutes)
2. You will also participate in a focus-group interview at the end of the session (~1 hour)
3. Four months after the end of the course, you will complete a writing task similar to an in-class writing assignment (~1 hour)
4. Your writing sample will be collected and analyzed.

Risks and Discomforts:

- We do not foresee any risks or discomfort from your participation in the research.

Benefits of the Research and Benefits to You:

- You may find the study useful for developing your writing proficiency by exploring your own writing process.
- Your contribution will also bring indirect benefits for future courses using computer and teachers" feedback and the advancement of understanding of the integration of computer feedback in the classroom.

Voluntary Participation and Withdrawal: Your participation in the study is completely voluntary and you may choose to stop participating at any time. Your decision not to volunteer, to stop participating, or to refuse to answer particular questions will not influence the nature of the ongoing relationship you may have with the researchers or study staff, or the nature of your relationship with York University either now, or in the future.

In the event you withdraw from the study, all associated data collected will be immediately destroyed wherever possible. Should you wish to withdraw after the study, you will have the option to also withdraw your data up until the analysis is complete.

Confidentiality:

- To assure confidentiality and privacy, fictitious names and alphanumeric codes will be used to assure your anonymity and confidentiality.
- Also, details may be omitted or changed that may make you easily identifiable.
- The key to the codes will be kept in a password protected file system away from the data set to prevent unauthorized access.
- You will have the opportunity to see a summary of the results.
- Unless you choose otherwise, all information you supply during the research will be held in confidence and unless you specifically indicate your consent, your name will not appear in any report or publication of the research.
- Recordings of your interview, questionnaires, and writing samples will be safely stored in a computer that is password protected and only the researcher will have access to this information.
- After the analysis is complete, all identifying information will be stripped and the data will be archived in a password protected zip file. Only the researcher will have access to this archived information. The data will be archived for future verification of the study and for further analysis using different analytical tools if needed.
- Confidentiality will be provided to the fullest extent possible by law.
- The data collected in this research project may be used – in an anonymized form - by members of the research team in subsequent research investigations exploring similar lines of inquiry.Such projects will still undergo an ethics review by the HPRC, our institutional REB.Any secondary use of anonymized data by the research team will be treated with the same degree of confidentiality and anonymity as in the original research

project.

- The researcher(s) acknowledge that the host of the online questionnaire (e.g., Qualtrix, Survey Monkey, etc.) may automatically collect student data without their knowledge (i.e., IP addresses). Although this information may be provided or made accessible to the researchers, it will not be used or saved without student"s consent on the researcher"s system. Further, because this project employs e-based collection techniques, data may be subject to access by third parties as a result of various security legislation now in place in many countries and thus *the confidentiality and privacy of data cannot be guaranteed during web-based transmission.*

Questions About the Research?If you have questions about the research in general or about your role in the study, please feel free to contact me at Johanathan_Woodworth@edu.yorku.ca or my supervisor, Khaled Barkaoui at KBarkaoui@edu.yorku.ca and/or (416) 736 2100 ext. 33209.You may also contact the Graduate Program in Education at gradprogram@edu.yorku.ca and/or 416-736-5018.

This research has received an ethics review and approval by the Delegated Ethics Review Committee, which is the delegated authority to review research ethics protocols by the Human students Review Sub-Committee, York University"s Ethics Review Board, and conforms to the standards of the Canadian Tri-Council Research Ethics guidelines.If you have any questions about this process, or about your rights as a student in the study, please contact the Sr. Manager & Policy Advisor for the Office of Research Ethics, 5th Floor, Kaneff Tower, York University (telephone 416-736-5914 or e-mail ore@yorku.ca).

Legal Rights and Signatures:

I _____ <<fill in student name here>>, consent to participate in Hybrid Feedback: The Efficacy of Combining Automated Writing Corrective Feedback and Teacher Feedback for Academic Writing Development in an English for Academic Purpose (EAP) Context conducted by Johanathan Woodworth.I have understood the nature of this project and wish to participate.I am not waiving any of my legal rights by signing this form.My signature below indicates my consent.

Signature _____ Date _____
student

Signature _____ Date _____
Principal Investigator

Additional consent

1. Audio recording

♦ I consent to the audio-recording of my interview(s).

I _____ <<insert students name>> consent to the use of images of me (including photographs, video and other moving images), my environment and property in the following ways (please check all that apply):

In academic articles	[] Yes	[] No
In print, digital and slide form	[] Yes	[] No
In academic presentations	[] Yes	[] No
In media	[] Yes	[] No
In thesis materials	[] Yes	[] No

Signature_____ Date_____

student

Informed Consent Form – Comparison group

Date:

Study Name: Hybrid Feedback: The Efficacy of Combining Automated Writing Corrective Feedback and Teacher Feedback for Academic Writing Development in an English for Academic Purpose (EAP) Context

Researcher name:
Johanathan Woodworth - PhD program in Education at York University (Principle investigator)
Johanathan_Woodworth@edu.yorku.ca

Khaled Barkaoui (Supervisor)
KBarkaoui@edu.yorku.ca
(416) 736 2100 ext. 33209

Purpose of the Research:

The purpose of the research is to investigate the effectiveness of teacher and machine feedback compared to teacher feedback only on student"s improvement in writing. The research will specifically examine language, content and organization of the writing.

The research will be conducted by collecting questionnaire, interview data and collecting your writing samples for analysis. The findings will be reported in a PhD dissertation.

What You Will Be Asked to Do in the Research:

1. Your writing sample will be collected and analyzed.
2. You will complete a questionnaire at the beginning of the semester (~10 minutes).

There will be no additional time commitment required from you for this research.

Risks and Discomforts:

- We do not foresee any risks or discomfort from your participation in the research.

Benefits of the Research and Benefits to You:

- You may find the study a useful for developing your writing proficiency by exploring your own writing process.
- Your contribution will also bring indirect benefits for future courses using computer and teachers" feedback and the advancement of understanding of the integration of computer feedback in the classroom.

Voluntary Participation and Withdrawal: Your participation in the study is completely voluntary and you may choose to stop participating at any time.Your decision not to volunteer, to stop participating, or to refuse to answer particular questions will not

influence the nature of the ongoing relationship you may have with the researchers or study staff, or the nature of your relationship with York University either now, or in the future.

In the event you withdraw from the study, all associated data collected will be immediately destroyed wherever possible. Should you wish to withdraw after the study, you will have the option to also withdraw your data up until the analysis is complete.

Confidentiality:

- To assure confidentiality and privacy, fictitious names and alphanumeric codes will be used to assure your anonymity and confidentiality.
- Also, details may be omitted or changed that may make you easily identifiable.
- The key to the codes will be kept in a password protected file system away from the data set to prevent unauthorized access.
- You will have the opportunity to see a summary of the results.
- Unless you choose otherwise, all information you supply during the research will be held in confidence and unless you specifically indicate your consent, your name will not appear in any report or publication of the research.
- Writing samples will be safely stored in a computer that is password protected and only the researcher will have access to this information.
- After the analysis is complete, all identifying information will be stripped and the data will be archived in a password protected zip file. Only the researcher will have access to this archived information. The data will be archived for future verification of the study and for further analysis using different analytical tools if needed.
- Confidentiality will be provided to the fullest extent possible by law.
- The data collected in this research project may be used – in an anonymized form - by members of the research team in subsequent research investigations exploring similar lines of inquiry.Such projects will still undergo an ethics review by the HPRC, our institutional REB.Any secondary use of anonymized data by the research team will be treated with the same degree of confidentiality and anonymity as in the original research project.

Questions About the Research?If you have questions about the research in general or about your role in the study, please feel free to contact me at Johanathan_Woodworth@edu.yorku.ca or my supervisor, Khaled Barkaoui at KBarkaoui@edu.yorku.ca and/or (416) 736 2100 ext. 33209.You may also contact the Graduate Program in Education at gradprogram@edu.yorku.ca and/or 416-736-5018.

This research has received an ethics review and approval by the Delegated Ethics Review Committee, which is the delegated authority to review research ethics protocols by the Human students Review Sub-Committee, York University"s Ethics Review Board, and conforms to the standards of the Canadian Tri-Council Research Ethics guidelines.If you have any questions about this process, or about your rights as a student in the study, please contact the Sr. Manager & Policy Advisor for the Office of Research Ethics, 5th Floor, Kaneff Tower, York University (telephone 416-736-5914 or e-mail ore@yorku.ca).

Legal Rights and Signatures:

I_____ <<fill in student name here>>, consent to participate in Hybrid Feedback: The Efficacy of Combining Automated Writing Corrective Feedback and Teacher Feedback for Academic Writing Development in an English for Academic Purpose (EAP) Context conducted by Johanathan Woodworth.I have understood the nature of this project and wish to participate.I am not waiving any of my legal rights by signing this form.My signature below indicates my consent.

Signature_____ Date_____
student

Signature_____ Date_____
Principal Investigator

Appendix I: Ethics Approval

YORK
UNIVERSITÉ
UNIVERSITY

OFFICE OF RESEARCH ETHICS (ORE)
5th Floor, Kaneff Tower

4700 Keele St.
Toronto ON
Canada M3J 1P3
Tel 416 736 5914
Fax 416 736-5512
www.research.yorku.ca

Certificate #:	**STU 2019-065**
Approval Period:	**06/21/19-06/21/20**

ETHICS APPROVAL

To: **Johnathan Woodworth**
Faculty of Education
Graduate Student
Johan.woodworth@gmail.com

From: Alison M. Collins-Mrakas, Sr. Manager and Policy Advisor, Research Ethics
(on behalf of Veronica Jamnik, Chair, Human Participants Review Committee)

Date: Friday June 21, 2019

Title: **Hybrid Feedback: The Efficacy of Combining Automated Writing Corrective Feedback and Teacher Feedback for Writing Development in an ESL Context**

Risk Level: ☒ Minimal Risk ☐ More than Minimal Risk

Level of Review: ☒ Delegated Review ☐ Full Committee Review

I am writing to inform you that this research project, **"Hybrid Feedback: The Efficacy of Combining Automated Writing Corrective Feedback and Teacher Feedback for Writing Development in an ESL Context"** has received ethics review and approval by the Human Participants Review Sub-Committee, York University's Ethics Review Board and conforms to the standards of the Canadian Tri-Council Research Ethics guidelines.

Note that approval is granted for one year. Ongoing research – research that extends beyond one year – must be renewed prior to the expiry date.

Any changes to the approved protocol must be reviewed and approved through the amendment process by submission of an amendment application to the HPRC prior to its implementation.

Any adverse or unanticipated events in the research should be reported to the Office of Research ethics (ore@yorku.ca) as soon as possible.

For further information on researcher responsibilities as it pertains to this approved research ethics protocol, please refer to the attached document, **"RESEARCH ETHICS: PROCEDURES to ENSURE ONGOING COMPLIANCE".**

Should you have any questions, please feel free to contact me at: 416-736-5914 or via email at: acollins@yorku.ca.

Yours sincerely,

Alison M. Collins-Mrakas M.Sc., LLM
Sr. Manager and Policy Advisor,
Office of Research Ethics

Appendix J: List of Indices for NLP Tools

List of indices for NLP Tools
From https://www.linguisticanalysistools.org/tools.html

TAACO: TOOL FOR THE AUTOMATIC ANALYSIS OF COHESION

Global Coreference Cohesion			
Index Name	**In text name**	**Index description**	**Denominator**
adjacent_overlap_all_para	adjacent paragraph overlap all lemmas	number of lemma types that occur at least once in the next paragraph	number of types in each paragraph (except the last paragraph)
adjacent_overlap_2_all_para	adjacent two-paragraph overlap all lemmas	number of lemma types that occur at least once in the next two paragraphs	number of types in each paragraph (except the last two paragraphs)
adjacent_overlap_noun_para	adjacent paragraph overlap noun lemmas	number of noun lemma types that occur at least once in the next paragraph	number of types in each paragraph (except the last paragraph)
adjacent_overlap_2_noun_para	adjacent two-paragraph overlap noun lemmas	number of noun lemma types that occur at least once in the next two paragraphs	number of types in each paragraph (except the last two paragraphs)
adjacent_overlap_verb_para	adjacent paragraph overlap verb lemmas	number of verb lemma types that occur at least once in the next paragraph	number of types in each paragraph (except the last paragraph)
adjacent_overlap_2_verb_para	adjacent two-paragraph overlap verb lemmas	number of verb lemma types that occur at least once in the next two paragraphs	number of types in each paragraph (except the last two paragraphs)
adjacent_overlap_argument_para	adjacent paragraph overlap noun and pronoun lemmas	number of noun and pronoun lemma types that occur at least once in the next paragraph	number of types in each paragraph (except the last paragraph)
adjacent_overlap_2_argument_para	adjacent two-paragraph overlap noun and pronoun lemmas	number of noun and pronoun lemma types that occur at least once in the next two paragraphs	number of types in each paragraph (except the last two paragraphs)
Global Conceptual Cohesion			
Index Name	In text name	Index description	Denominator
syn_overlap_para_noun	adjacent paragraph overlap noun synonyms (paragraph normed)	number of noun lemma types that occur at least once in the next paragraph (inclusive of synonyms of each noun lemma type)	number of paragraphs in text (except last paragraph)

syn_overlap_para_verb	adjacent paragraph overlap verb synonyms (paragraph normed)	number of verb lemma types that occur at least once in the next paragraph (inclusive of synonyms of each verb lemma type)	number of paragraphs in text (except last paragraph)
Local Coreference Cohesion			
Index Name	In text name	Index description	Denominator
adjacent_overlap_all_sent	adjacent sentence overlap all lemmas	number of lemma types that occur at least once in the next sentence	number of types in each sentence (except the last sentence)
adjacent_overlap_2_all_sent	adjacent two-sentence overlap all lemmas	number of lemma types that occur at least once in the next two sentences	number of types in each sentence (except the last two sentences)
adjacent_overlap_noun_sent	adjacent sentence overlap noun lemmas	number of noun lemma types that occur at least once in the next sentence	number of types in each sentence (except the last sentence)
adjacent_overlap_2_noun_sent	adjacent two-sentence overlap noun lemmas	number of noun lemma types that occur at least once in the next two sentences	number of types in each sentence (except the last two sentences)
adjacent_overlap_verb_sent	adjacent sentence overlap verb lemmas	number of verb lemma types that occur at least once in the next sentence	number of types in each sentence (except the last sentence)
adjacent_overlap_2_verb_sent	adjacent two-sentence overlap verb lemmas	number of verb lemma types that occur at least once in the next two sentences	number of types in each sentence (except the last two sentences)
adjacent_overlap_argument_sent	adjacent sentence overlap noun and pronoun lemmas	number of noun and pronoun lemma types that occur at least once in the next sentence	number of types in each sentence (except the last sentence)
adjacent_overlap_2_argument_sent	adjacent two-sentence overlap noun and pronoun lemmas	number of noun and pronoun lemma types that occur at least once in the next two sentences	number of types in each sentence (except the last two sentences)
Local Conceptual Cohesion			
Index Name	In text name	Index description	Denominator
syn_overlap_sent_noun	adjacent sentence overlap noun synonyms (sentence normed)	number of noun lemma types that occur at least once in the next sentence (inclusive of synonyms of each noun lemma type)	number of sentences in text (except last sentence)
syn_overlap_sent_verb	adjacent sentence overlap verb synonyms (sentence normed)	number of verb lemma types that occur at least once in the next sentence (inclusive of synonyms of each verb lemma type)	number of sentences in text (except last sentence)
Givenness for Text Cohesion			

Index Name	In text name	Calculation Method	
repeated_content_lemmas	repeated content lemmas	number of repeated content lemmas divided by number of words	
pronoun_density	pronoun density	number of third person pronouns divided by number of words	
Connectives			
Index Name	In text name	Description	Denominator
all_connective	all connectives	number of all connectives	number of words in text

TAALES: TOOL FOR THE AUTOMATIC ANALYSIS OF LEXICAL SOPHISTICATION

Lexical Depth						
Index Name	In Text Name	Category	Description	Numerator/ equation	Denominator	Types of Words
All_AWL_Normed	Academic Word List All	Academic Language	Normed Count	Number of AWL words in text	Number of words in text	N/A
COCA_academic_bi_T	COCA Academic Bigram Association Strength (T)	Ngram Association Strength	Mean T Association Strength Score	Sum of T scores	number of bigrams in text with T scores	All Words
COCA_academic_tri_T	COCA Academic Trigram Unigram to Bigram Association Strength (T)	Ngram Association Strength	Mean T Association Strength Score (item 1 = first word, item 2 = following bigram)	Sum of T scores	number of trigrams in text with T scores	All Words
COCA_academic_tri_2_T	COCA Academic Trigram Bigram to Unigram Association Strength (T)	Ngram Association Strength	Mean T Association Strength Score (item 1 = first bigram, item 2 = remaining word)	Sum of T scores	number of trigrams in text with T scores	All Words
hyper_noun_Sav_Pav	Hypernymy Nouns (Sense Mean, Path Mean)	Semantic Network	Average hypernymy score for nouns (average for all senses, all paths)	Sum of hypernymy scores	Number of words in text with hypernymy scores	Nouns
hyper_verb_Sav_Pav	Hypernymy Verbs (Sense Mean, Path Mean)	Semantic Network	Average hypernymy score for verbs (average for all senses, all paths)	Sum of hypernymy scores	Number of words in text with hypernymy scores	Verbs

content_poly	Polysemy CW	Semantic Network	Average number of senses for content words	Sum of polysemy scores	Number of words in text with polysemy score	Nouns, Verbs, Adjectives, and Adverbs
poly_noun	Polysemy Nouns	Semantic Network	Average number of senses for nouns	Sum of polysemy scores	Number of words in text with polysemy score	Nouns
poly_verb	Polysemy Verbs	Semantic Network	Average number of senses for verbs	Sum of polysemy scores	Number of words in text with polysemy score	Verbs
poly_adj	Polysemy Adjectives	Semantic Network	Average number of senses for adjectives	Sum of polysemy scores	Number of words in text with polysemy score	Adjectives
poly_adv	Polysemy Adverbs	Semantic Network	Average number of senses for adverbs	Sum of polysemy scores	Number of words in text with polysemy score	Adverbs
Lexical Frequency						
Index Name	In Text Name	Category	Description	Numerator/ equation	Denominator	Types of Words
COCA_Academic_Frequency_AW	COCA Academic Frequency AW	Word Frequency	Mean Frequency Score	Sum of frequency scores	number of words in text with frequency score	All Words
COCA_Academic_Frequency_CW	COCA Academic Frequency CW	Word Frequency	Mean Frequency Score	Sum of frequency scores	number of words in text with frequency score	Content Words
COCA_Academic_Frequency_FW	COCA Academic Frequency FW	Word Frequency	Mean Frequency Score	Sum of frequency scores	number of words in text with frequency score	Function Words
COCA_Academic_Bigram_Frequency	COCA Academic Bigram Frequency	Ngram Frequency	Mean bigram frequency score	Sum of bigram frequency scores	number of bigrams in text with frequency score	All Words
COCA_Academic_Trigram_Frequency	COCA Academic Trigram Frequency	Ngram Frequency	Mean trigram frequency score	Sum of trigram frequency scores	number of trigrams in text with frequency score	All Words

Lexical Range						
Index Name	In Text Name	Category	Description	Numer-ator/ equation	Denominator	Types of Words
COCA_Academic_Range_AW	COCA Academic Range AW	Word Range	Mean Range (number of documents that a word occurs in) score	Sum of range scores	number of words in text with range score	All Words
COCA_Academic_Range_CW	COCA Academic Range CW	Word Range	Mean Range (number of documents that a word occurs in) score	Sum of range scores	number of words in text with range score	Content Words
COCA_Academic_Range_FW	COCA Academic Range FW	Word Range	Mean Range (number of documents that a word occurs in) score	Sum of range scores	number of words in text with range score	Function Words
COCA_Academic_Bigram_Range	COCA Academic Bigram Range	Ngram Range	Mean bigram range score	Sum of range scores	number of bi-grams in text with range score	All Words
COCA_Academic_Trigram_Range	COCA Academic Trigram Range	Ngram Range	Mean trigram range score	Sum of range scores	number of trigrams in text with range score	All Words

TAASSC: TOOL FOR THE AUTOMATIC ANALYSIS OF SYNTACTIC SOPHISTICATION AND COMPLEXITY

Index Name	SCA Name	Source	Index Type	Description	Numerator
Length of production unit					
MLS	Syntactic Complexity Analyzer	Unit Length	mean length of sentence	number of words in text	number of sentences in text
MLT	Syntactic Complexity Analyzer	Unit Length	mean length of T-unit	number of words in text	number of T-units in text
MLC	Syntactic Complexity Analyzer	Unit Length	mean length of clause	number of words in text	number of clauses in text
Sentence complexity					
C/S	Syntactic Complexity Analyzer	Clausal	clauses per sentence	number of clauses in text	number of sentences in text
Amount of subordination					
C/T	Syntactic Complexity Analyzer	Clausal	clauses per T-unit	number of clauses in text	number of T-units in text
CT/T	Syntactic Complexity Analyzer	Clausal	complex T-unit ratio	number of complex T-units in text	number of T-units in text

DC/C	Syntactic Complexity Analyzer	Clausal	dependent clauses per clause	number of dependent clauses in text	number of clauses in text
DC/T	Syntactic Complexity Analyzer	Clausal	dependent clauses per T-unit	number of dependent clauses in text	number of T-units in text
Amount of coordination					
CP/C	Syntactic Complexity Analyzer	Clausal	coordinate phrases per clause	number of coordinate phrases	number of clauses in text
CP/T	Syntactic Complexity Analyzer	Clausal	coordinate phrases per T-unit	number of coordinate phrases	number of T-units in text
T/S	Syntactic Complexity Analyzer	Clausal	T-units per sentence	number of T-units in text	number of sentences in text
Degree of phrasal sophistication					
CN/C	Syntactic Complexity Analyzer	Clausal	complex nominals per clause	number of complex nominals	number of clauses in text
CN/T	Syntactic Complexity Analyzer	Clausal	complex nominals per T-unit	number of complex nominals	number of T-units in text
VP/T	Syntactic Complexity Analyzer	Clausal	verb phrases per T-unit	number of verb phrases in text	number of T-units in text

Noun Phrase Complexity	
Index Name	In Text Name
det_all_nominal_deps_struct	determiners per nominal
amod_all_nominal_deps_struct	adjectival modifiers per nominal
prep_all_nominal_deps_struct	prepositions per nominal
poss_all_nominal_deps_struct	possessives per nominal
vmod_all_nominal_deps_struct	verbal modifiers per nominal
nn_all_nominal_deps_struct	nouns as a nominal dependent per nominal
rcmod_all_nominal_deps_struct	relative clause modifiers per nominal
advmod_all_nominal_deps_struct	(non-clausal) adverbial modifiers per nominal
conj_and_all_nominal_deps_struct	conjunction "and" as a nominal dependent per nominal
conj_or_all_nominal_deps_struct	conjunction "or" as a nominal dependent per nominal

Appendix K: Descriptive Statistics for the Writing Process Questionnaire by Group and Time Point

Descriptive Statistics for the Writing Process Questionnaire by Group and Time Point

	Experimental				**Comparison**				**Increase in Scores**	
	Pretest		**Posttest**		**Pretest**		**Posttest**		**Comp.**	**Exp.**
	M	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**		
Conceptualization Phase										
1. I read the whole prompt carefully	3.24	0.44	3.53	0.51	3.13	0.50	3.38	0.5	0.29	0.25
2. I thought about how well I understood the task requirements	3.24	0.66	3.41	0.62	3.13	0.62	3.38	0.62	0.17	0.25
3. I thought about what I know about the topic	3.12	0.6	3.35	0.61	3.19	0.54	3.75	0.45	0.23	0.56
4. I thought about what I know about the genre	2.94	0.43	3.29	0.59	2.94	0.57	3.13	0.72	0.35	0.19
5. I thought about the purpose of the task	2.71	0.77	3.29	0.47	2.81	0.54	3.38	0.62	0.58	0.57
6. I thought about what I might need to write to make my essay relevant and adequate to the task	2.82	0.88	3.29	0.59	3.19	0.54	3.63	0.5	0.47	0.44
7. I thought about the intended reader of my essay and their expectations	2.47	0.8	2.94	0.56	3.19	0.40	3.44	0.51	0.47	0.25
15. I thought about the appropriate words to express my ideas in the first draft	2.94	0.83	3.29	0.59	3.19	0.66	3.25	0.58	0.35	0.06
23. I changed my writing plan (e.g., structure and content) in the first draft	2.29	0.99	3.06	0.83	3.00	0.89	3.06	0.68	0.77	0.06
24. I checked that the content is relevant and revised accordingly in the first draft	2.41	0.8	3.29	0.47	3.13	0.62	3.44	0.63	0.88	0.31
Generating Ideas Phase										
8. I thought about or jotted down ideas which are relevant to the task/topic	3.12	0.6	3.35	0.61	3.19	0.54	3.31	0.48	0.23	0.12
9. I prioritized my ideas based on the task requirements	2.88	0.86	3.24	0.66	3.31	0.6	3.44	0.63	0.36	0.13
10. I linked my ideas based on the task requirements	3.12	0.7	3.41	0.62	3.31	0.7	3.38	0.62	0.29	0.07

11. I worked out how my ideas relate to each other, e.g., main ideas or examples	2.82	0.73	3.53	0.51	3.31	0.48	3.38	0.50	0.71	0.07
22. I re-read the task instructions/ prompts in the first draft	2.65	0.93	3.29	0.69	3.19	0.66	3.31	0.70	0.64	0.12
Organizing Ideas Phase										
12. I thought about the structure of my essay	2.94	0.9	3.41	0.62	3.56	0.51	3.69	0.48	0.47	0.13
13. I removed some ideas I planned to write because they did not fit the structure of my essay	2.94	0.75	3.18	0.73	3.00	0.89	3.19	0.83	0.24	0.19
14. I re-read the prompt/task instructions	2.53	1.01	3.06	0.90	3.25	0.77	3.38	0.81	0.53	0.13
19. I thought about how to make my ideas persuasive to the reader in the first draft	2.65	0.70	3.06	0.75	3.25	0.45	3.50	0.52	0.41	0.25
21. I brain-stormed main and supporting ideas in the first draft	2.94	0.56	3.35	0.61	3.50	0.63	3.50	0.63	0.41	0.00
Generating Texts Phase										
16. I thought about the correct sentence structures to express my ideas in the first draft	2.65	0.79	3.35	0.61	3.38	0.62	3.56	0.63	0.70	0.18
17. I thought about the correct grammar to express my ideas in the first draft	2.76	0.9	3.41	0.80	3.56	0.63	3.50	0.63	0.65	-0.06
18. I thought about how to connect my ideas smoothly in the whole essay in the first draft	2.94	0.83	3.24	0.75	3.19	0.40	3.25	0.45	0.30	0.06
20. I organized my sentences and paragraphs in a logical order in the first draft	2.59	0.62	3.24	0.56	3.19	0.54	3.25	0.45	0.65	0.06
Monitoring and Revising at High-Level Phase										
25. I checked that my essay is well-organized and revised accordingly in the first draft	2.65	0.70	3.35	0.70	3.00	0.73	3.06	0.77	0.70	0.06
26. I checked that my essay is coherent and revised accordingly in the first draft	2.59	0.87	3.18	0.81	3.19	0.54	3.31	0.70	0.59	0.12
27. I checked that I include my own viewpoint on the topic and revised accordingly in the first draft	2.59	0.80	3.29	0.59	3.19	0.66	3.44	0.73	0.70	0.25

28. I checked the possible effect of my essay on the intended reader and revised accordingly in the first draft	2.41	0.71	3.12	0.60	3.31	0.60	3.38	0.62	0.71	0.07
29. I checked the accuracy and range of sentence structures and revised accordingly in the first draft	2.53	0.72	3.41	0.62	2.94	0.77	3.19	0.54	0.88	0.25
33. I checked that my essay was well organized and revised accordingly after receiving feedback	2.00	1.46	3.29	0.69	2.50	0.82	2.88	0.81	1.29	0.38
34. I checked that my essay was coherent and revised accordingly after receiving feedback	2.00	1.41	3.29	0.69	2.94	0.68	3.06	0.68	1.29	0.12
35. I checked that I include my own viewpoint on the topic and revised accordingly after receiving feedback	2.06	1.52	3.24	0.56	2.81	0.66	3.06	0.85	1.18	0.25
36. I checked the possible effect of my essay on the intended reader and revised accordingly after receiving feedback	1.76	1.35	3.18	0.64	2.94	0.68	3.13	0.81	1.42	0.19
37. I checked the accuracy and range of the sentence structures and revised accordingly after receiving feedback	1.88	1.36	3.41	0.62	2.75	0.68	3.31	0.70	1.53	0.56
Monitoring and Revising at Low-Level Phase										
30. I checked the grammar (e.g., part of speech and tenses) and revised accordingly in the first draft	2.82	0.73	3.29	0.85	3.06	0.77	3.25	0.68	0.47	0.19
31. I checked the appropriateness and range of vocabulary and revised accordingly in the first draft	1.65	0.79	3.35	0.49	2.75	1.00	2.88	0.81	1.70	0.13
32. I usually write multiple drafts after receiving feedback	2.06	1.48	3.29	0.69	2.81	0.75	2.69	0.87	1.23	-0.12
38. I checked the grammar (e.g., part of speech and tenses) and revised accordingly after receiving feedback	1.88	1.45	3.35	0.70	2.75	0.58	2.94	0.44	1.47	0.19
39. I checked the appropriateness and range of vocabulary and revised accordingly after receiving feedback	1.76	1.30	3.24	0.83	2.63	0.72	3.06	0.77	1.48	0.43

Appendix L: Cronbach"s Alpha results Descriptive Statistics for the Writing Process Questionnaire

Cronbach"s Alpha results Descriptive Statistics for the Writing Process Questionnaire

	Cronbach's Alpha	N of Items
Conceptualization Phase Pretest	0.741	10
Conceptualization Phase Posttest	0.727	10
Generating Ideas Phase Pretest	0.743	5
Generating Ideas Phase Posttest	0.673	5
Organizing Ideas Phase Pretest	0.656	5
Organizing Ideas Phase Posttest	0.63	5
Generating Texts Phase Pretest	0.822	4
Generating Texts Phase Posttest	0.836	4
Monitoring and Revising at High-Level Phase Pre-test	0.913	10
Monitoring and Revising at High-Level Phase Posttest	0.886	10
Monitoring and Revising at Low-Level Phase Pre-test	0.877	5
Monitoring and Revising at Low-Level Phase Posttest	0.625	5

N=33