Effects of task repetition and feedback on integrated listening-to-write performances and perceptions of English for Academic Purposes university students

John Bandman

This thesis is submitted for the degree of

Doctor of Philosophy (PhD) in Linguistics

Department: School of Social Sciences

Discipline: Linguistics and English Language

Lancaster University, United Kingdom

October 27, 2025

This thesis is my own work and has not been submitted in substantially the same form for the award of a higher degree elsewhere.

Abstract

The use of tasks prevails in second language (L2) teaching in many contexts nowadays. A particular pedagogic technique that is thereby sometimes used is that of task repetition. Task repetition studies have been conducted to explore the effects repetition has on L2 production in terms of the complexity, accuracy and fluency (CAF) of the task performance. The majority of studies' findings suggest that as students focus their attention on learning a task's requirements, CAF dimensions compete with one another. This aligns with Skehan's Trade-Off Hypothesis (1998a), which posits that at the initial task performance, the L2 learner focuses on meaning, but when they repeat the task, they focus more on improving form (grammar). Most studies have explored effects on oral performances, while fewer are available on written performances, typically elicited with independent writing tasks. In many L2 teaching and assessment contexts, however, writing constructs have broadened to include integrated tasks. To the best of my knowledge, very few studies have yet investigated repetition effects of such tasks, including listening-to-write tasks. A first aim of this study was therefore to investigate the effect that task repetition has on a listening-to-write task in terms of the CAF of the writing performance. Additionally, given the listening input, it also explored effects of knowledge summary and knowledge transfer from the listening into writing. Furthermore, as feedback is considered an important pedagogic tool, the study also examined the effect of feedback on task repetition. Finally, the study wanted to explore learners' perceptions of this kind of task and repetition.

Data were collected from 64 upper-intermediate university students in an English for Academic Purposes (EAP) programme in the United States, randomly split into two groups: feedback and no-feedback. Each student completed the task, then repeated the task one week after the first performance, then two weeks after the second performance. The feedback group received coded metalinguistic and holistic feedback within two days after their first two performances. To gauge student perceptions, a task perception questionnaire was administered after the third performance.

Mixed-between-within ANOVA analyses revealed significant main effects of task repetition for both groups for the majority of CAF measures, knowledge summary and knowledge transfer. At the second performance, for both groups, there was evidence of competition among some of the CAF dimensions, i.e., a trade-off, which partially aligns with trade-off effects revealed in earlier studies. More specifically in the second performance, for the feedback group, there was a partial trade-off between accuracy and

some complexity measures where there were no significant improvements; for the no-feedback group, there was a trade-off between accuracy and some complexity and fluency measures where there were significant declines. In the third performance, for the feedback group, there was less competition between the CAF dimensions, which aligns with Sample and Michel's (2014) finding that trade-off effects disappear in the third performance. For the no-feedback group, trade-off effects did not disappear.

No significant main effects of feedback were found between the feedback and no-feedback groups' written performances. However, there were interaction effects between feedback condition and repetition for some of the CAF measures, thus feedback had positive pedagogic effects with the help of repetition even though there were no main effects in isolation. In terms of perceptions, regardless of whether students had received feedback, they overall held positive views about this task and about repeating tasks. This study suggests that using and repeating integrated listening-to-write tasks in upper-intermediate L2 learning is worthwhile even if the students do not receive feedback.

Acknowledgments

My journey from inception to completion of this thesis would not have been possible or as enriching a learning experience without the help and support of various people who have guided me in immeasurable ways.

First and foremost, I thank Prof. Tineke Brunfaut, my supervisor, for her heartfelt dedication to guiding me, and for her detailed feedback, support, continued recommendations for success, and selfless heartfelt desire to be there to set her students up for success. I also thank Elaine Heron who has always been there whenever I sought logistical guidance at the university.

I thank Prof. Luke Harding, Dr. John Pill, Prof. Patrick Rebuschat, and Prof. Tineke Brunfaut for leading research groups in the Department of Linguistics and English Language. I also am grateful that they made it possible to participate in the groups both on campus and remotely. Such interaction, most notably during the pandemic and beyond, has enabled me to share ideas and receive feedback from peers and faculty plus offer support for others where I could. I extend my special thanks to my fellow PhD students for their time and support.

I have a long list of friends and professionals to thank for their support, educationally, practically, morally, or a combination. In particular, I thank Dr. Christine Rosalia, Dr. Alexander Jay, Ronda Drakeford, John Mancuso, Iris Bucchino, Charles Bordogna, Dr. Pierre LaGuerre, Jovani Tablezo, Dr. Barry Freeman, Tomer Zilkha, Aaron Morrisey, and Simon Vaz.

I thank my entire family and friends for their love, encouragement, and support, and for giving me further strength to strive for excellence during my PhD journey. I am forever grateful that you were always there.

I am also very thankful for many people who were not there during my PhD journey, mainly those who were very understanding when I needed to excuse myself from some engagements. This extra time enabled me to meet deadlines at various stages of my studies, including coursework, data collection, analysis, thesis development, review, revisions, and completion. As I complete this rigorous yet rewarding journey, I look forward to re-engaging in many of those plans.

Table of Contents

List of Figures

List of Tables

1 Introduction

Task repetition is a pedagogic technique used by second language teachers in many contexts, including in English for Academic Purposes (EAP). Researchers and educators alike have therefore been keen to uncover the way repetition works – what repetition does and how it affects language learning. Researchers and educators have also been interested in the extent to which a combination of both repetition and feedback has a positive impact on language production. So far, most research in this area has focused on oral language production; a much more limited number of studies has looked at ways in which repetition and feedback can improve written language production. Furthermore, the type of tasks explored so far primarily concerns independent skills tasks such as speaking-only or writing-only. Authentic language uses, however, often involve multiple language skills – a combination of at least two skills from reading, writing, listening and/or speaking. Language education has also come to recognise this by extending the typical 'four skills' approaches to additionally focusing on integrated language uses. To date, however, comparatively less research is available on this, and to the best of my knowledge there are very few studies available so far on the effect of repetition of, and feedback on, integrated tasks (e.g., Kim & Kim, 2017).

1.1 Aims statement

The aim of this study is to help close this gap by looking at the impact that repeating the same integrated listening-to-write task at three intervals, in an EAP course, has on aspects of the task performance. More specifically, the study aims to investigate effects on the complexity, accuracy, and fluency (CAF) of students' written performances, and on the knowledge summary and knowledge transfer from the listening input into the writing. Brief definitions of these terms are as follows, and will be elaborated on in Chapter 2:

Complexity: the more difficult and challenging use of language.

Accuracy: the closeness of adherence to a type of norm in a language, i.e., linguistic form/grammatical structure.

Fluency: the ease and smoothness of speech or writing.

Knowledge summary: a summary, in a condensed format, of the main content points only from the input; knowledge summary can include the retelling (written or spoken) of recently learned key content from the input (listening or reading comprehension).

Knowledge transfer: communication (written or spoken) of the use or application of recently learned information from the input (e.g., reading or listening) into similar or new contexts (which can also be an integrated task).

Additionally, this study looks into the effect that feedback has on the CAF, knowledge summary and knowledge transfer characteristics of the integrated writing performances, and the interaction between repetition and feedback. Further, the study probes into opinions that students have about task repetition and about the listening-to-write task explored in my study

1.2 Brief description of this study's EAP setting

The study reported in this thesis was conducted at a public university in a major city in the northeastern United States, where I worked as a lecturer. Students aged 18 and above from 155 different countries study at this university, where there is a wide range of majors that they can pursue. More specifically, the study was set in the university's EAP language immersion programme that offers intensive instruction to students who are accepted into the university, but whose first language is not English and whose placement examination results suggest that they need to take further EAP courses before they are ready to take credit-bearing courses. Each EAP course takes place for a 15-week semester, and students are allowed to take a maximum of three semesters of EAP classes. Once students complete their studies in the EAP programme, they move on to their major.

In the EAP immersion programme, students are in class 25 hours per week with the same teacher for the entire semester, and the course prepares students for academic college writing, reading, speaking, listening and study skills. In terms of writing, part of the course involves intensive writing such that students go to the computer lab every day for at least an hour to do research, listen to videos and recorded materials, and write essays. In terms of integrated listening-to-write, for example, the students listen to newscasts, write about the information they heard, and then respond to essay prompts.

1.3 Rationale for the study

1.3.1 Independent skill tasks vs. real-world language uses

Since the spread of the communicative language teaching approach from the 1970s onwards, independent-skill tasks, which involve only one of the four language skills at a time (listening, reading, speaking, writing), have been widely used in language education, including independent writing tasks in teaching and assessing L2 student writing. The communicative approach promotes communicative activities that revolve around the controlled delivery of pre-selected linguistic items whereby students eventually practice in free production (Samuda & Bygate, 2008). In independent writing tasks, students write based on their background knowledge and prior experience; thus, their success in completing an independent task risks being influenced by prior familiarity (or lack thereof) with the topic more so than by language

skill. In other words, students may perform better when they are familiar with the given topic than when they are not. Similarly, if the students are not familiar with the topic, it may result in an underestimation of their actual writing ability. When educators use such an independent task for assessment as a pedagogic technique, it can be problematic because a variable irrelevant to writing constructs (i.e. background knowledge) may affect students' task performance and interpretations of any scores allocated to it. Such construct-irrelevant variance (Messick, 1989, p. 44) poses a threat to a task's construct validity.

Additionally, an increasing number of researchers have argued that independent writing tasks have limitations in terms of authenticity and that they constitute construct-underrepresentation in many contexts, including academic ones. Real-world communication, i.e., academic, social, and professional, often requires the use of at least two language skills in integration, not in isolation, for example, reading-to-write and listening-to-speak. For instance, real-world demands in professional contexts often involve professionals to rely on information that they have heard or read to then make decisions, develop responses, or produce written reports. Weigle (2004) states, "academic writing is rarely done in isolation, but is virtually always done in response to source texts" (p. 30). Cumming (2013) and Shin and Ewert (2015) have similarly pointed out that most real-life academic writing involves the integration of what one has read (e.g. journal articles) and/or heard (e.g. lecture and seminar input) into one's writing. Indeed, Yang (2009) verified that in academic contexts in terms of writing, students work with source materials to identify, synthesize, connect, and manipulate data in their writing. Therefore, it is evident that an independent task may not always align with academic language use. Instead, language teachers and researchers in academic contexts have increasingly supported the use of what have been called integrated writing tasks in the language classroom, as these have been argued to increase authenticity and validity (Cumming, Kantor, Power, Santos, & Taylor, 2000; Weigle, 2004). As Yang (2009) summarizes, an important motivation for the inclusion of integrated writing tasks in present-day language teaching and assessment as a pedagogic technique in academic settings is that these tasks are "reflective of the real use of language that occurs in academic contexts" (p. 3). Indeed, Lewkowicz (1997) supported that integrated tasks are intended to resemble the language situations that students often experience in academic contexts.

An integrated task requires the student to combine at least two language skills – in traditional terms, a receptive/input skill (listening and/or reading) and a productive/output skill (speaking and/or writing). The integration of the various language skills means that the language output is based on the language input, and that aspects of the input are integrated into

the output, as opposed to two or more separate skills operating independently of one another. The authenticity justifications mentioned above have been used for the adoption of integrated tasks in high-stakes EAP language tests (Cumming, 2014; Cumming et al., 2000). For example, reading-to-write tasks are currently included in the Canadian Academic English Language (CAEL) Assessment, the Pearson Test of English (PTE) Academic, and the Test of English as a Foreign Language Internet-based (TOEFL® iBT). Outside of academic settings, they are also found in numerous standardized and institution-based general English proficiency tests such as the Georgia State Test of English Proficiency (GSTEP) and the General English Proficiency Test (GEPT).

1.3.2 Integrated skills tasks in this study's EAP setting

Following a similar rationale, the university language immersion program where I work continually employs integrated language skills activities and assigns integrated coursework to students. The demands of the program's upper-intermediate level EAP classes in particular, which are the focus of this study, can be considered high; students in these classes are at the level of going on to academic degree program study, so they must be prepared to cope with integrated tasks.

In the immersion program, some of the integrated language tasks that the students complete incorporate reading-to-write, reading-to-speak, and listening-to-speak. Listening-to-write is also sometimes required, but comparatively less frequently than reading-to-write or listening-to-speak. The fact that this has been less frequently used has made this an attractive task type to research in this study. An example of a listening-to-write task used in the program is one that requires the students to watch and listen to a talk such as TED Talk, and then perform a writing activity pertaining to the input. The use of this newscast activity shows that listening-to-write already exists in the EAP program, thus making this type of skill integration relevant for the present empirical research. It is thought that writing tasks with listening input enhance students' abilities to process and synthesize spoken language in real-time. This skill integration enhances which applies their understanding of the input to incorporate that into their language output. This integrated language skill is necessary for completing everyday academic tasks as well as in the workplace, for example, note-taking. While listening-to-write tasks share their reliance on oral input processing with the more frequently used listening-to-speak tasks, it is argued that the written output as opposed to the spoken output makes the task type distinct. For example, apart from the obvious differences between speaking and writing,

listening-to-write also allows students to practice and employ the ability to reflect on what they communicate about the input with the ability to revise their output.

A strength of using a listening-to-write task in our EAP program is that it gives us information on the progress that students make in listening, writing, and in the ability to integrate the two skills. Using just a series of independent listening tasks in isolation and writing tasks in isolation could not only be more resource- and time-consuming, but would also not shed light on students' ability to integrate these skills and thus miss out on representing an important feature of many real-life language uses. While integrated tasks may require more cognitive skills than independent tasks, i.e. students need to process and refer back to information from the input – a skill that independent tasks do not require (e.g., Brown et al., 2005; Cumming et al., 2006) – this does reflect the demands of language use in many contexts. The ability to successfully apply skill integration is necessary for academic and professional environments, which are the target language use domains of our EAP students.

Also, even though in many parts of the world the formats and modes of education and workspaces are changing, i.e. a shift from face-to-face meetings to hybrid classrooms and remote workspaces, listening-to-write activities still form part and parcel of these 'new' environments, e.g. following a remote lecture or meeting, taking notes, and transforming these into written texts such as reports or summaries. Thus, listening-to-write skill development remains valid in our EAP program from this perspective too.

1.3.3 Task-based language teaching

The listening-to-write task focused on in this study aligns with conceptualizations of tasks in the field of Task-Based Language Teaching (TBLT) in that it is meaning-focused rather than primarily a vocabulary or grammar drill. TBLT is sometimes seen as a teaching method by itself (e.g., Kumaravadivelu, 2006), while others view it as a perspective within the Communicative Language teaching (CLT) framework that emphasizes tasks as opposed to its being a distinct method of instruction by itself (e.g., Brown, 2007, Willis & Willis, 2007). In these views, for example, TBLT is characterized as a method where meaning is primary and connected to real-life. These characteristics can help make TBLT motivating because it is relevant to learners' real-life needs (Skehan, 1989, 1998a). The aim of TBLT is for the activities to be relevant to the students' experiences and that they give them practice in situations where they may find themselves in the future. These aims are also shared by the language immersion program where this study took place.

1.3.4 Task repetition

A regular practice in the immersion program, in terms of writing development work, is that students are required to do at least two revisions of their written work ('essays' – a broad denominator for a range of writing tasks). After the students complete the revisions, the final drafts are usually ready to be placed on display in the university hallways. Some of the writing tasks are independent skill tasks, others are integrated ones, with reading-to-write ones prevailing over listening-to-write ones to date.

Results from a wide body of research on language student performances reveal that a key outcome of repetition is that it helps students develop their second language proficiency and improve their performances (Muhammadpour et al., 2023; Hsu, 2019; Amiryousefi, 2016; Azizzadeh & Dobakhti, 2015; Kim & Tracy-Ventura, 2013; Jung, 2013; Ahmadian & Tavakoli, 2011). This suggests that as students become more familiar with the task, the repetition helps the students not only retain the language input but the familiarity with the task helps them improve their ability to complete the task while using the newly learned language. Taking it to the next step, repetition helps students successfully strengthen their skills as they practice using the new content and language in different contexts.

Within TBLT more specifically, the use of task repetition has been motivated by Skehan's Trade-Off Hypothesis (1998a). Formerly known as the Limited Attentional Capacity Model, trade-off proposes that when a student performs a task for the first time, their capacity is limited in that they need to allocate their concentration between the process and the task demand itself. As the student is completing the task, the CAF dimensions compete with one another as the student learns the demands of the task completion.

There is some variation in terms of where the trade-off occurs as well as the specific CAF components that are stronger than the others. For example, several studies have found that with repetition, the trade-off is between accuracy and complexity, notably complexity (e.g., lexical sophistication) being stronger than accuracy (e.g., grammatical structures) during the trade-off (e.g., Gass et al., 1999; Sample & Michel, 2014). During a repetition, some of the CAF dimensions benefit from the repetition, but it is not always the same dimensions across comparative studies that benefit or compete with one another. For example, results from Ahmadian and Tavakoli's (2011) study on the combined effect of oral task repetition and planning conditions revealed that complexity and fluency benefited from the repetition while, at the same time, there were no significant changes to accuracy.

Next is Hsu's (2019) study where there were three groups: one group that performed task repetition and post-transcription, one group that repeated the task but did not perform

post-transcription, and the control group that did not perform task repetition or post-transcription. Unlike the results from Ahmadian and Tavakoli's study, results from Hsu's (2019) comparative study on the combined effect of oral task repetition and post-transcribing revealed that for the repetition group, there were significant gains in accuracy, but no significant differences between the two groups in complexity and fluency.

Most repetition studies have looked into one repetition; fewer studies have looked into two repetitions. Several studies have found that trade-off effects disappear at the third performance, i.e., second repetition (e.g., Sample and Michel, 2014; Bygate, 2009; Hawkes, 2009, 2012). For example, Sample and Michel (2014) conducted an investigative oral task repetition study on the effects on language output in terms of CAF with two repetitions (Time1-original, Time2-first repetition, Time3-second repetition). The results revealed that at Time2, there was a trade-off between complexity (complex sentences and lexical sophistication) and accuracy (grammatical structures), at which time complexity was stronger than accuracy. However, at Time3, the trade-off effects disappeared.

While a body of empirical research has tested the trade-off hypothesis by investigating task repetition effects, most studies have focused on repetition of oral performances (e.g., Bygate, 1996, 2001; Gass et al., 1999; Ahmadian & Tavakoli, 2011; Kim & Tracy-Ventura, 2013; Sample & Michel, 2014; Bui et al., 2019; Hsu, 2019; Hassanzadeh-Taleshi et al., 2023; Muhammadpour et al., 2023). Only a small number of studies have investigated task repetition effects for writing (e.g., Jung, 2013; Azizzadeh & Dobakhti, 2015; Amiryousefi, 2016). Within the latter set of work, to the best of my knowledge, very few studies have investigated repetition of integrated writing (see section 2.3.5), exploring mainly reading-to-write, reading-listening-write, and listening-to-speak, but very few so far have investigated this for listening-to-write. My study therefore attempts to address this gap by applying the trade-off hypothesis to a new task type. Theoretically, my study hypothesises there are similar trade-off effects (e.g., accuracy-complexity) that result from the repetition of an integrated listening-to-write task as that which originally applied for oral performances, thereby contributing to this existing theory.

1.3.5 Feedback

Feedback plays a major role for the students in the EAP immersion program. For one, it makes them aware of their progress, mistakes, and areas for improvement. Another reason why the program emphasizes the importance of providing students with feedback is motivational, i.e., it can give them a sense of recognition for their progress and efforts (of course, assuming good-quality and supportive feedback). For the instructor, because the EAP classes run five days per

week with the same instructor, providing feedback can help them gain clearer insights into each student's progress and areas of improvement. This can also enable instructors to encourage students to maintain their momentum and achieve their next learning goals. In order for feedback to have positive impacts on student learning, it should be specific, targeted, and timely (Allman, 2019).

According to Allman (2019), one main reason for the higher effectiveness of specific feedback (vs. general feedback) is that it enables students to immediately recognise their areas for improvement. From the educator's side, the feedback focus should be directly relevant to the task's goals. The feedback should be targeted, clear, and easy for students to recognise what they need to do to improve in their future work. Feedback should also be prompt so that students can review their work while they are still on topic.

Some of the feedback research was conducted in the context of task repetition (e.g., Nguyen et al, 2023; Kim & Kim, 2017), with some on independent/integrated speaking tasks, and some on independent/integrated writing tasks. For example, Nguyen et al.'s (2023) study focused on an independent speaking task and investigated the impact that task repetition had in conjunction with post-task-teacher-corrected transcribing (feedback) on speaking performances in terms of CAF. Kim and Kim's (2017) study presents a rare example of an integrated writing task in this area of research. It investigated the impact that feedback had on the written performances of repeated integrated reading-to-write tasks. The results of both these task repetition studies suggested that there were favorable gains in student work as a result of feedback as a main effect. I am interested in finding out whether feedback, specifically in the context of task repetition, has an effect on integrated listening-to-write task performances, which to the best of my knowledge is not something that has been explored before.

1.4 Structure of the thesis

This thesis is composed of six chapters. Following this introduction chapter is the literature review (Chapter 2). In the latter chapter, first the theoretical framing of this study is introduced. Then literature on TBLT is reviewed in detail as it relates to facets of this study. Then CAF is defined, followed by a review of CAF measures used in previous empirical research. After that, task repetition is discussed as a technique, through a review of empirical studies that have looked at various oral and written task repetitions in terms of CAF, knowledge summary and knowledge transfer. Research on feedback and on perceptions of task repetition is also presented, followed by a review of research on integrated tasks, knowledge

summary and knowledge transfer. Finally, informed by the prior literature, I present my hypotheses and research questions.

Chapter 3 describes the research methodology of this study. First, the overall research design, setting, and participants are described. Next, the data collection methods, measures, and research instruments, comprising a background questionnaire, listening-to-write task, and task perception questionnaires are explained. After that, the data collection and data analyses of the study are presented. Finally, the analysis of the task perception questionnaire is discussed.

In Chapter 4, the findings are presented. Descriptive and comparative statistics are provided for all CAF measures and knowledge summary and transfer. Findings for student perceptions about this task, the listening-to-write integration, and about task repetition are also presented. Feedback condition group comparisons are made for all of the measures.

In Chapter 5, all these findings are then discussed in association with the research questions and the literature. This chapter also shows the extent to which the findings align with my hypotheses. Further, it shows how my study contributes to and furthers our knowledge on repetition, feedback and integrated tasks in L2 teaching.

Chapter 6 concludes this study by first summarizing the study and its main findings. In addition, the contributions of my study and its pedagogical implications are outlined. In conclusion, the limitations of this study and recommendations for future research are indicated.

2 Literature Review

In this chapter, I first introduce and discuss TBLT in section 2.2. I then explain the definitions for CAF in section 2.3.1, followed by a presentation of CAF measures, along with examples of empirical studies that have used the measures. After that, I delve into task repetition as used in pedagogy after which I detail empirical studies that have incorporated oral task repetition and CAF (section 2.3.4) and written task repetition and CAF (section 2.3.5). Then I review the literature on feedback in task repetition (section 2.4), perceptions of task repetition (section 2.5), integrated tasks (section 2.6), and knowledge summary and transfer (section 2.7). Finally, I present my research questions and hypotheses (section 2.8).

2.1 Theoretical Framing

I position the task repetition element of my research within the existing theory in TBLT, which emerged in the 1980s. More specifically, it is motivated by Skehan's Trade-Off Hypothesis (1998a, 2009). The trade-off hypothesis was previously known as the Limited Attentional Capacity Model (Skehan, 1998a; Skehan & Foster, 2001). This hypothesis, developed in the context of oral production tasks, proposes that "due to capacity limitations, speakers must divide their attentional resources between all the processes a task requires…If various task demands exceed the available resources, [then complexity, accuracy, and fluency] compete with each other" (Sample & Michel, 2014, p. 27). This division of attentional resources occurs because learners must make choices of where they will achieve their gains in their task completion output at the expense of other aspects of their task completion output.

The trade-off hypothesis claims that during the first time completing a task, learner focus is primarily on meaning, but when a task is repeated, the learner is able to shift their focus increasingly on language form as they perform their task. Some research that investigated task performances after one repetition has shown that the trade-off is mostly between accuracy and complexity (Foster & Skehan, 1996; Skehan, 1998b). There is also some research on two task repetitions instead of one. For example, Sample and Michel (2014) finding suggests that a trade-off occurs in that "initial performances that benefit in one dimension come at the expense of another; by the third performance, trade-off effects disappear…that with growing task-familiarity students are able to focus their attention on all three CAF dimensions simultaneously" (p. 23). Skehan (2009), however, does not claim an automatic requirement for symmetrical, simultaneous, significant performance increases and decreases for there to be a trade-off, i.e., although such trade-offs are likely, improvements do

not always have to be accompanied by losses in language production performances to occur in order for it to be called a trade-off; it can be a lack of gain or a decrease.

In my research, I attempt to adopt parts of the Limited Attentional Capacity Model for written language performances as a result of task repetition and to determine whether results from the present study affirm my hypothesis that such repetition will have similar results in written tasks as the trade-off hypothesis had stated for oral tasks.

2.2 Task-Based Language Teaching (TBLT)

As its name suggests, task-based learning involves students learning language through completing tasks; TBLT's main tenet is that the key element in language lesson plans, curriculum and assessments must be a task (Bygate, 2016; Ellis, 2009; Samuda & Bygate, 2008). As Richards and Rodgers (2001) stated, TBLT uses tasks as the core planning and instructional unit.

An important feature of TBLT is that the tasks are intended to reflect real-world activities such as replying to e-mails, making a business call, attending a lecture, participating in a business meeting, etc. (Long, 2015). Indeed, Lightbown and Spada (1993) defined TBLT as instruction in which class activities are tasks similar to those that learners might engage in outside of the L2 classroom. Long (2015) summed up TBLT as "an approach to course design, implementation, and evaluation intended to meet the communicative needs of diverse groups of learners" (p. 5).

2.2.1 Definition of task in TBLT

Within TBLT, the basic requirement of a task is that it is a classroom activity "that requires students to use language in a meaningful communicative way to achieve an outcome" (Faez & Tavakoli, 2019, p. 7). In reviewing the literature, researchers offer additional meanings and descriptions that can be ambiguous to practitioners who attempt to distinguish the term 'task' from an activity, drill or exercise (Crookes, 1986; Long, 1985,). Ellis (2003) pointed out that task definitions vary in terms of scope of the task, authenticity, required language skills, cognitive process, perspective from which the task is viewed, and the task's outcome. In fact, Bygate et al. (2001) indicate that definitions of tasks are usually context-free, which is part of the reason why narrowing down a definition is problematic, notably in that a task will have various meanings in different contexts of use. Therefore, I draw from various researchers' explanations of a task.

Some researchers provided a broad definition of a task that does not make the language focus explicit. For example, Long (1985) defined task as "a piece of work undertaken for oneself or for others, freely or for some reward" (p. 85), possibly to deduce that Long took the language context for granted. In contrast, other researchers, for example, Richards et al. (1985) and Nunan (1989) included only activities that involved language as tasks. Nunan (1989) stated that a task is "a piece of classroom work which involves learners in comprehending, producing or interacting in the target language while their attention is principally focused on meaning rather than form" (p. 10).

Additionally, later researchers supported outcome as a key underpinning of what defines a task. Willis (1996) stressed that a task is "a goal-oriented activity in which learners use language to achieve a real outcome" (p. 53). Willis supported that a task is any activity that always includes the target language used by the learner to achieve an outcome through communicating the newly learned L2. According to Willis, tasks are goal oriented with very specific outcomes such that the target language is used in a meaningful way to finish the task, require completion, and to relate to real world outcomes as opposed to producing specific forms. This description of a task was later supported by Bygate et al. (2001); they defined a task as "an activity which requires to use language, with an emphasis on meaning to attain an objective" (p.11). In later work, Bygate (2016) defined tasks more specifically as activities where learners use language pragmatically to do things with the overriding aim of learning language" (p. 381).

Building on a task's definition to focus on meaning and outcome, several researchers helped streamline the criteria of a task, with some variation of what constitutes a task. According to Skehan (1998a, 1998b), a task has four key characteristics which differentiate tasks from exercises. Similar to the task definitions introduced above, Skehan (1998a, 1998b) posited that in a task: (1) the meaning is primary; (2) a communication problem such as an information gap needs to be solved; (3) the task has some sort of relationship to real-life activities that people do; and (4) the task is assessed in terms of its outcome, not in terms of linguistic features. In his view, evaluators are interested not only in the way language has been used in the task performance but also whether the task's communicative purpose is met. Thereby, there must be a relationship between the activity that arises from the task and the way that language is used in the real world (Skehan, 1998a, 1998b).

Widdowson (1998), on the other hand, was critical of Skehan's (1998a, 1998b) task criteria, arguing that the "criteria do not in themselves distinguish the linguistic exercise and the communication task" (p. 328). He argued that a task and an exercise differ regarding the

kind of meaning, goal and outcome. Widdowson argued that an exercise is planned based on a need to develop the linguistic skills as a prerequisite for a communicative activity, while a task is based on assumptions that linguistic skills are developed through the communicative activity.

As far as developing linguistic skills is concerned, later on, Richards and Rodgers (2001) explained that tasks are not form-focused. Richards and Rodger's criteria for a task are: (1) use of language previously learned or introduced at pre-task stage; (2) non-language-driven outcomes; (3) relevance to learner needs; (4) allowing to reflect on language use; and (5) dependency on communication and interaction skills. Like Richards and Rodgers, D. Willis and S. Willis (2001) stated that tasks are used for communicative purposes to attain an outcome. While these claims suggest that the achievement of a communicative goal is subservient to an instructional agenda, further empirical investigations need to be conducted to validate it.

Another common theme from the task definitions and descriptions mentioned above is the major role that tasks play, which is that improving language learning hangs on the practice of language use as primarily a way to make meaning. According to Ellis (2003), a task is a type of workplan involving primary focus on meaning and real-world language use, and incorporation of the language skills while drawing on cognitive processes to fill an information gap. To expand on making meaning-driven outcomes, Ellis (2000, 2003) draws from Prabhu (1987) who identifies three types of tasks: (1) information gap where students must convey information with one another; (2) reasoning gap that requires learners to create new information from existing information in order to draw inferences and make deductions; and (3) opinion gap that is based on students' commentary useful for discussions. These show that a task provides a clearly stated communicative outcome, thus illustrating different types of intended communicative outcomes to convey meaning. In addition to meaning-making or the need for an outcome, Ellis (2009) later on adds to Prabhu's earlier constitutes of a task that learners should need to rely on their own linguistic or non-linguistic knowledge to complete an activity to show their understanding of the meaning of the task. In sum, most researchers emphasize the use of a task in language learning as a "tool for achieving a communicative outcome rather than language itself being the object to be studied although some focus-on-form is necessary" (Ellis & Shintani, 2013, p.136). While various researchers offer their definitions of tasks, the following are common criteria:

- primary focus on meaning
- dependency on communication and interaction skills.

- gap where students must convey information
- learners should need to rely on their own linguistic or non-linguistic knowledge
- target language is used in a meaningful way to finish the task as opposed to producing specific forms

2.2.2 Uses of tasks in TBLT

TBLT has been developed and introduced by language educators and second language acquisition researchers as a response to the teacher-fronted and form-focused methods of language teaching (Van den Branden, 2006). Since the 1980s, the literature on tasks has become even more focused on establishing an empirical basis for using tasks as the primary organizational component in L2 instruction and establishing accountable learning outcomes associated with this approach (Skehan 1996; Robinson 2011). Nunan (2004) advocated that tasks are "an activity or action which is carried out as a result of a process to understand a language [such as] drawing a map, performing a command…." (p. 7). A few years earlier, Nunan (1999) supported that tasks help promote L2 acquisition in that the learners gain further exposure to the language by rehearsing it to prepare for practical uses of the L2. Some examples of practical goals achieved from completing a task include describing something from a text or lecture, finding main points, recalling facts and details, and providing a description (Zuniga, 2016). Butler (2011) argued that the use of tasks is encouraged and incorporated as required components in English language curricula. In fact, Abdelaty (2023) argued that "…tasks simulate real-world situations, promoting natural language use and fostering the development of all language skills, including speaking, listening, reading, and writing" (p. 238), thus constituting opportunities for learners to make many meaningful real-world connections to their language learning.

Within the field of second language acquisition (SLA), there has been a growing interest in the use of pedagogical tasks in classroom contexts (Loewen, 2015 & Khezrlou, 2019), and TBLT's popularity rests on the following: (1) Tasks provide an ideal second language (L2) processing environment by presenting meaningful language use opportunities as well as a timely reactive focus on linguistic problems in context as they arise during task performance (Long, 2015) and (2) it is a learner-centered approach since it attends to learners' needs as the key part of development and success of language programs (Lee, 2018). This learner-centered approach is also about learners' needs as well as the active role that the learner needs to fulfill. Further, the teacher should not be instructing from A to Z regarding the completion of the tasks. These key strengths of TBLT have led to a considerable amount of

research concerning task design, implementation, planning and repetition variables with the purpose of optimizing TBLT (Bygate, 2001; Ellis, 2016; Long, 2015, Khezrlou, 2019).

The focus in TBLT is to acquire language through task performance rather than intentionally in preparation for doing linguistic-driven tasks. Tasks are designed to provide learners with a communicative need for using language and with an opportunity for negotiation of understanding based on their own L2 resources. In turn, learners notice and fill gaps in their linguistic resources to develop the interactive competence necessary to complete tasks and communicate effectively in the target language. Tasks are also argued to motivate learners either through connections to their real-world needs (Long, 2015; Lambert & Oliver, 2020) or through the interest that educators have for the learners (Ellis et al., 2019; Lambert & Oliver, 2020).

TBLT seeks to provide learners with a natural context for language use: "As learners work to complete a task, they have abundant opportunity to interact. Such interaction is thought to facilitate language acquisition as learners have to understand each other and to express their own meaning" (Larsen-Freeman 2000, p.144). However, TBLT is not a fixed mechanism but rather implemented through a variety of tasks, for example, reciprocal, non-reciprocal, target and pedagogical tasks (Abbassian & Chenabi, 2016, p. 3).

2.3 Task repetition and CAF

In this section, I discuss CAF as well as task repetition. This present study uses an integrated listening-to-write task; it is important to note that the writing is analysed by CAF while the listening comprehension is analysed by knowledge summary and transfer. (The latter is discussed in 2.7).

One pertinent line of research in the empirical literature on TBLT concerns the repetition of tasks and its effect on learning. Task repetition, defined as "repetition of a given configuration of purposes, and a set of content information" (Bygate, 2018, p. 2), is one of the methodology options in task-based language learning. Task repetition can involve repeating the exact same or partially different tasks at various time intervals (Bygate & Samuda, 2005), yet still there is usually a repetition of the familiarity that the students should have with the form and content used to be prepared to fulfill the task (Bygate, 2006).

Task repetition has been argued to be effective in improving aspects of speaking ability such as fluency (Ahmadian, 2011; Bygate, 2001; Gass et al., 1999; Li & Rogers, 2021; Lynch & Maclean, 2000). The common rationale for task repetition is that repeating tasks positively improves a learner's L2 output with respect to complexity, accuracy and fluency (CAF)

(Bygate, 1996, 2001; Gass, et al., 1999; Lynch & McLean, 2001). The effect of task repetition has consequently often been explored through the measures of CAF. These, in fact, have been key measures discussed in second language acquisition research since the 1980s when distinctions were made between fluent and accurate use of language (Housen & Kuiken, 2009; Skehan, 1998a), and have become commonly used in research studies as dependent variables. Results from the majority of task repetition studies have revealed that task repetition improves CAF in speaking, but the results vary in terms of the specific CAF components where the improvements occurred (e.g. Ahmadian & Tavakoli*,* 2011; Bygate, 1996, 2001; Khezrlou, 2021).

Within the body of task repetition research that has mainly examined the effect of task repetition on L2 CAF, studies conducted to date have mostly focused on two key lines of research, oral performance in the same task (e.g. Boers, 2014; Bygate, 1996; Gass et al., 1999; Lambert et al., 2017; Muhammadpour et al., 2023; Thai & Boers, 2016; Wang, 2014) and oral performance in a new task (e.g. Kim & Tracy-Ventura, 2013; Van de Guchte et al., 2016). Oral performance research has yielded consistent findings in terms of positive effects of both same and new task repetitions on L2 learners' fluency but varying findings in terms of complexity and accuracy. Some studies have observed improvement in all three CAF measures (e.g. Ahmadian & Tavakoli, 2011; Kim & Tracey-Ventura, 2013). Other studies, however, report large increases in complexity and slight increases in accuracy (Gass et al., 1999) or increases in fluency and accuracy, but not complexity (Bygate, 1996; Lynch & McLean, 2000), or increases in fluency and complexity, but not accuracy (Bygate, 2001). Further studies, examining multiple repetitions in a single class period, have found immediate positive impacts on fluency (Thai & Boers, 2016; Lynch & McLean, 2000) and slight gains in accuracy (Lynch & McLean, 2000).

It should be noted, however, that different studies have examined the impact of exact task repetition (same task, same content), procedural repetition (same task, new content), and/or content repetition (same content, different task) for speaking tasks across days or weeks during L2 development, operationalized as CAF. So, task repetition is not always operationalized in the same manner. Also, many of the studies that investigated CAF used different set of CAF measures, or rather they described the meanings more generally (Housen et al, 2012). Additionally, different CAF measures are used in different studies. These might explain some of the differences in findings between studies. As a result, “this limits the interpretation and comparability of CAF findings and may also explain why the CAF literature has produced many inconsistent findings” (Housen et al, 2012, p. 4). These thereby might

make the generalizability of some of the CAF studies difficult to deduce.

In order to understand how L2 learners allocate their attention in the context of task repetition, it is important to know what scholars have argued happens during the first and second task performances. During the first task performance, L2 learners process meaning during what has been described as a 'task rehearsal' phase (Bygate, 1996). Following Levelt and Speaking's (1989) model, learners prioritize the conceptualization stage over the formulation and articulation stages. However, during second or subsequent repetitions of the same task, learners tend to skip the conceptualization stage (Bygate & Samuda, 2005) because they already know the meaning and input communication content, thereby allocating their attention and monitoring-based resources to the formulation and articulation stages. In fact, the repetition of the same or similar tasks may support learners to build from "what they have already done in order to buy time not only to do mental work on what they are about to communicate but also to access and (re)formulate words and grammatical structures more efficiently, effectively, and accurately" (Ahmadian, 2012, p.380).

There is, however, more limited research available on the language skill of writing in this context. When it comes to writing, it is thought that for many beginner writers who are given a specific task for the first time, using their transcription and oral fluency skills necessary to decode words' meanings challenges their abilities to free up their working memory capacity required for critical thinking skills (Kellogg, 2001; Olive & Kellogg, 2002; Torrance & Galbraith, 2005). Amiryousefi (2016) indicates for writing: "…the existence of more time compared with speaking and the visibility of the text produced, learners have more time and opportunity to pay attention to form and meaning simultaneously and to involve more active monitoring" (p. 1054). This helps encourage learners to go over their writing even without repetition because, arguably, this can also simply be that they had one hour instead of thirty minutes, so not necessarily as a result of repetition, hence why comparative studies can help confirm this. This suggests that with task repetition, the benefit is learners have more processing space available to formulate better language structures to accomplish the task in the repetition performance. Potentially, this is why they can produce a more complex, accurate, and fluent L2 output with variable rates.

In the next two sections, I will provide some research definitions of CAF along with measures used, then I will explain the importance of task repetition as a pedagogical technique. Subsequent to that, I will review empirical studies that have examined oral task repetition and written task repetition and their effects on CAF (sections 2.3.4 and 2.3.5, respectively).

2.3.1 Definitions of CAF

CAF has been used to measure language learning progress for the last few decades (Housen & Kuiken, 2009; Crave, 2017). It is necessary to clearly identify the terms and the means of measurements in order to compare one study to another, thus a clarification of specific CAF meanings and measures is central in order to accurately draw conclusions. Each research study is done with a specific purpose and population. Therefore, researchers define the terms and select measures suitable for their studies. The variation in CAF definitions makes the concepts of complex, accurate and fluent "polysemic, vague and with different meanings in the ordinary language versus technical domains" (Pallotti, 2021, p. 205).

CAF research has been conducted since the 1970s. At that time, researchers looked at the learners' first language as a way to measure language proficiency "in an objective, quantitative and verifiable way" (Housen et al, 2012, p.2). In this case, learners drew on insights of their L1 to help them learn an L2. Researchers had established concepts and measures for CAF based on L1 research, and they were trying to apply them to L2 learning research contexts. Second language pedagogy researchers measured fluency and accuracy but did not measure complexity when they explored communicative second language use (Brumfit, 1979).

In the 1990s, Skehan (1996, 1998a, 1998b) brought together the three CAF dimensions, i.e. complexity, accuracy, and fluency. This was also when the CAF definitions became more fine-tuned (Housen et al, 2012). **Complexity** is generally defined in the literature as the more difficult and challenging use of language. **Accuracy** is about the ability to produce language that reflects a type of norm in the language. An example is "a learner's capacity to handle whatever level of inter-language complexity s/he has currently attained" (Skehan, 1996, p. 46), which includes the language produced by the learner accurately reflecting the target language. **Fluency** is the ease and smoothness of speech or writing (Michel, 2017; Chambers, 1997).

However, as mentioned above, researchers need to know what they are comparing to draw conclusions when reading across studies. Many of the studies that investigated CAF used different set of CAF measures, or rather they described the meanings more generally (Housen et al, 2012). Consequently, it has become challenging to get to more generalizable findings because each study used measures in a different way. Therefore, it made it difficult to directly compare findings across studies because methodologically and conceptually, they are not all the same and not all based on the same foundation. As a result, "this limits the interpretation and comparability of CAF findings and may also explain why the CAF literature has produced many inconsistent findings" (Housen et al, 2012, p. 4), thereby making the generalizability of the CAF studies difficult to deduce.

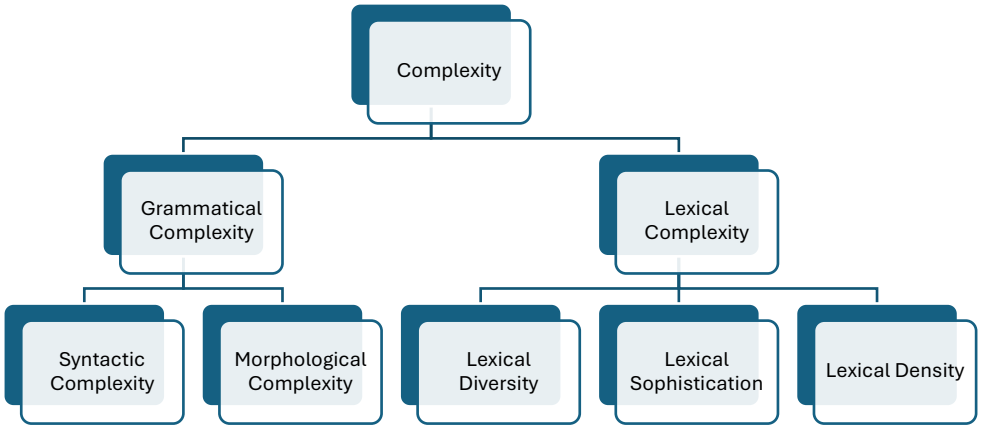
Next, I further define the CAF terms after which I identify and discuss the measures that were used in previous empirical studies.

2.3.1.1 Defining complexity

Several definitions support the general description of complexity as the more difficult and challenging use of language. For example, Ellis (2023) described complexity as "the extent to which the language produced…is elaborated and varied" (p. 240). Housen et al. (2012) defined complexity as elaboration by stating the extent to which a learner's performance includes the use of "a wide and varied range of sophisticated structures and vocabulary in the L2" (p. 2).

Often, different subtypes of complexity are distinguished depending on the particular element of language focused on, e.g. morphology, lexis, etc. Figure 2.1 visualizes a chart that identifies key categories of complexity, which are introduced next.

Figure 2.1. *Breakdown of complexity*



Grammatical complexity

Many L2 tests contain items that focus on sentence structure and grammatical features, representing various levels of difficulty. For example, on lower proficiency tests, it would be rather simple to select a correct response whereas more difficult tests would require learners to demonstrate a higher level of grammatical proficiency, i.e., by responding to open-ended prompts such as writing an essay or delivering a speech.

Grammatical complexity constructs are a critical L2 writing feature that has been widely used to assess the quality of language development (Lan et al, 2019). Grammatical complexity has been defined in various fields within the study of linguistics. It is important to note that grammatical complexity is also about level of difficulty and variety, not just structures in and of themselves. In psycholinguistics, for example, it is a complexity measured by the amount of time a learner needs to understand a grammatical structure (Newmeyer & Preston, 2014). In second language acquisition (SLA), it refers to grammatical structures

stemming from linguistic features and relationships within the phrases (Pallotti, 2015), as well as variety/level of difficulty. In L2 writing, as learners progress in language proficiency, they typically produce more advanced grammatical structures (Thirakunkovit & Rhee, 2021; Biber et al, 2011).

Syntax and morphology are the two specific domains that learners need to improve their grammatical complexity to enhance their constructive communication and language development. **Syntax** is the set of rules that govern ways words and phrases are formed to create complex sentences and phrases, for example, subordinating conjunctions like 'although' to show relationships between a phrase and the rest of the clause. **Morphology** is the set of internal word structures and their meaningful parts in relation to the rules to form words, for example, adding an 's' to form pluralization or "ly' to form an adverb. In sum, morphology provides words based on a language's rules, and syntax gathers the words to organize into a proper order to construct phrases and sentences. These two domains are interdependent and necessary for a learner to achieve grammatical complexity.

Syntactic complexity and morphological complexity are subcategories of grammatical complexity, and now I explain these terms.

Syntactic complexity

In terms of syntactic complexity (grammatical sophistication demonstrated in L2 production), different types of complexity have been described in the literature. Ortega (2003) defines syntactic complexity as "the range of forms that surface in language production and the degree of sophistication of such forms" (p. 492). Later on, Pallotti (2015), differentiated three types of syntactic complexity: First, structural (proper arrangement of texts and linguistics systems pertaining to their relational patterns); second, cognitive (pertaining to the learner's processing associated with the structures); and third, developmental ("the order in which language structures emerge and are mastered in SLA") p. 118). Syntactic complexity aligns with various degrees of linguistic elaborateness, diversity, and formality, thereby directly pertinent to academic language (Biber, 2016; Ortega, 2015 & 2003).

Morphological complexity

Morphological complexity is the range of internal structure of words (morphological forms) with a text. Grammatical proficiency depends on morphological knowledge to effectively communicate, and the more that a word's internal parts (morphemes) that are present, then the more changes to the word's dictionary forms, for example, verb tense, pluralization, etc., thus the more complex the language becomes (Brezine & Pallotti, 2016).

There are many irregularities in languages. Developing proficiency in the more morphologically complex words can confuse many learners. Learners with a stronger understanding of the grammar in their L1, as well as similarities in grammar between their L1 and L2, might find these irregularities easier to understand. Brezine and Pallotti (2016) exemplify morphologically complex languages with formal word parts expressing grammatical or word-form functions. With such variation among different languages, "the relationships among different parts of the system cannot be straightforwardly derived from a small set of systematic rules" (p. 100), thus intensifying morphological complexity. For example, morphology, like syntax, contains rules for placing together elements within any language but with varying ways to express the same function, for instance, adding an 'e' at the end of a noun in the French language to show the word's feminine gender.

Lexical complexity

Lexical complexity comprises the size, variation, and quality of a learner's vocabulary usage. It is the range and sophistication of vocabulary produced in written or spoken language (Wolfe-Quintero et al., 1998). Like grammatical complexity, lexical complexity is also a construct with subcategories. Measures of lexical complexity tap into three aspects of lexical performance: diversity, sophistication, and density (Skehan, 2023; Bulté et al., 2008; Bulté & Housen, 2012). Vocabulary usage that includes more words that are unique and infrequent in a learner's text is an indication of that learner's higher proficiency and text quality (Friginal et al., 2014; Kormos, 2011; Wolfe-Quintero et al., 1998; Zenker & Kyle, 2021).

Lexical diversity is about how different words are from each other in a text. **Lexical sophistication** is "the proportion of relatively unusual or advanced words in a learner's text…rather than just general everyday vocabulary" (Read, 2000, p. 203). This proportion compares the percentage use of simpler high-frequency words to the percentage of more advanced low-frequency words, and this ratio can be used to predict further writing quality and level of formality (Qin & Wen, 2007; Zhang et al., 2022). **Lexical density** is the proportion of lexical words (those with independent definitions) to all words within the text (Ure,1971). Ure (1971) and Halliday (1989) indicated that earlier research studies that compared lexical density between written and spoken texts suggest that written texts contained more lexical words.

2.3.1.2 Defining accuracy

Accuracy, in short, is the level of correctness of the learner's use of the language in terms of grammar, vocabulary, and pronunciation. Several researchers offer further descriptions of accuracy's definition. Wolfe-Quintero et al. (1998) described accuracy as "the conformity to

certain norms" (p. 4). Similarly, Pallotti (2009) described accuracy as "the simplest and most internally coherent construct, referring to the degree of conformity to norms" (p. 592). According to Housen and Kuiken (2009), if a learner falls short on accuracy, it means that the "… L2 performance deviated from a norm" (p. 4). Housen et al. (2012) identified accuracy as the most straightforward construct within the CAF triad.

2.3.1.3 Defining fluency

Fluency is the efficiency, naturalness, and smoothness of language production used to create coherent ideas. For speech, fluency is generally defined as the learner's ability to demonstrate the extent to which they can produce speech or text in terms of ease and confidence like a highly proficient speaker (Chambers, 1997; Lennon, 1990; Housen et al., 2012). It is the ease with which a learner can speak easily and quickly with minimal need to pause. For writing, fluency is a learner's ability to write with fluidity and with confidence, therefore about the ease of writing instead of speaking. What sets spoken and written fluency apart is the speaking versus writing skill.

A number of researchers provided broader descriptions of fluency. Wolfe-Quintero et al. (1998) described fluency as the comfort level that the speaker or writer has at retrieving the L2 while using the language. Additionally, fluency requires speakers "to draw on their memory-based system, assessing and deploying ready-made chunks of language" (Ellis, 2003, p. 113). Plakans et al. (2019) stated that fluency "captures a communicator's ability to think, compose, and deliver a language in real-time" (p. 164). With specific reference to writing, fluency has been the most persistently used measure to distinguish writing across proficiency/score levels for both independent and integrated tasks (Cumming et al., 2006; Ferris, 1994; Gebril & Plakans, 2013; Watanabe, 2001).

2.3.2 Measures of CAF

In this section, I provide examples of ways in which CAF has been measured and empirical studies that have adopted those CAF measures that I define. Whenever possible, I prioritized examples of empirical studies of task repetition to exemplify where the CAF measures were adopted. Where I did not provide examples of task repetition in writing, I gave other examples, for example, oral task repetition or other language studies that used CAF measures. I also state what was found for the specific measure under scrutiny.

Measuring complexity

Complexity has been measured by educators and researchers using either holistic (subjective) or quantitative (objective) measures, or both (Bulté & Housen, 2012). Complexity is often viewed as the most controversial of the CAF measures (Craven, 2017; Polio, 2011; Michel, 2017). Polio (2011) stressed that both syntactic and lexical features of complexity can be measured in L2 production. Later on, Ortega (2015) suggested that for different proficiency levels, some aspects of complexity might be more relevant for some learning levels, therefore there are ranges of complexity that can be measured such as sentential, clausal and lexical. Like Ortega, Pallotti (2021) recommended various measures of complexity, and Pallotti recommended that the complexity measures represent syntactic, morphological and lexical features of language production. The assumption behind these measures is learners will produce more elaborate and complex language as their proficiency develops.

2.3.2.1 Syntactic complexity measures

In terms of assessing **syntactic** complexity, a number of measures have been put forward in the literature. Wolfe-Quintero et al. (1998) discussed research that measured the ratios of grammatical features by dividing the total number of that specific grammar structure by the total number of t-units (see Table 2-1). T-units and clauses are key terms used for some of the accuracy measures. Polio (1997) defined a t-unit as "an independent clause and all its dependent clauses" (p. 139). Polio defined a clause as "an overt subject and a finite verb" (p. 139). A frequently used measure is the number of clauses per t-unit (C/T) to measure subordination, an aspect of sentential complexity, i.e., an increased number of clauses in proportion to t-units is an increase in sentential complexity. Other measures frequently used in previous studies include average sentence length, number of dependent clauses per clause (DC/C), and number of dependent clauses per T-unit (DC/T). Norris and Ortega (1979) referred to these measures as "variety, sophistication, and acquisition of forms produced" (p. 562).

Table 2-1 shows key measures and their definitions, then I briefly describe their uses in a couple of empirical studies and the findings for those measures.

Table 2-1. *Syntactic complexity measures*

Construct	Measure	Examples of studies that have used this measure
Sentential complexity	Average sentence length (mean # of	Nitta & Baba, 2018; Barrot and Agdeppa, 2021; Larsson & Kaatari, 2020; Manchon et al, 2023; Teng & Huang, 2021; Kormos, 2011

words per sentence)	
Clauses per T-unit (C/T) (mean # of clauses per T-unit)	Azizzadeh & Dobakhti, 2015; Jung, 2013; Lu, 2011; Barrot and Agdeppa, 2021; Ghahderijani, 2021; Kizil, 2023; Indrarathne, 2013; Larsson & Kaatari, 2020; Rathi, 2020; Rokoszewska, 2022; Tai, 2015; Teng & Huang, 2021
Dependent clause per clause (DC/C) (mean # of dependent clauses per clause)	Lu, 2011; Sang & Zou, 2023; Manchon et al., 2023; Ghahderijani, 2021; Phuoc, 2022; Ping-Ju, 2020; Teng & Huang, 2021
Dependent clause per T-unit (DC/T) (mean # of dependent clauses per T-unit)	Lu, 2011; Larsson & Kaatari, 2020; Phuoc, 2022; Rokoszewska, 2022; Teng & Huang, 2021

2.3.2.1.1 Average sentence length

Average sentence length is measured by establishing the mean number of words per sentence. Following Ortega (2003), Nitta and Baba (2018) used average sentence length as one of the syntactic complexity measures in a writing task repetition study, and these measures were calculated by Coh-Metrix mechanical detection. In their study, Nitta and Baba were looking at complexity and fluency effects of repeating a writing task over one year. The data were based on weekly 10-minute timed self-reflection L2 compositions that twenty-six L2 university students at a Japanese university completed. The students repeated the same task two weeks in a row, then wrote about new topics in the subsequent sets of two-week periods when they were given new sets of the same task to repeat during those two weeks, thus following the same procedure but new topics in alternating weeks. In earlier studies, it was observed that results based on this measure varied based on L2 proficiency. In Nitta and Baba's (2018) study, results showed that there were incremental increases in average sentence length for the first half of the year and then larger increases in the second half.

Another study that used this measure, which was calculated by the L2SCA measuring tool, is Barrot and Agdeppa's (2021). They analysed over 5000 essays from the International Corpus Network of Asian Learners of English (ICNALE), aiming to deduce interactions

between CAF measures and the four L2 proficiency levels that were investigated. Part of the goal was to glean insights into ways researchers can select various measures of language proficiency. Their results showed that average sentence length reached statistical differences when comparing this measure between proficiency levels in that there were steadily higher averages as the proficiency levels progressed.

2.3.2.1.2 Clauses per T-unit (C/T)

C/T is a frequently-used measure of the amount of subordination. C/T is measured by establishing the mean number of clauses per T-unit. According to Crossley and McNamara (2014), C/T is used as a measure because of its capability to predict future syntax and writing proficiency. At the same time, when raters do manual counts, they must be trained to identify clauses that are coordinated versus subordinated. Dependent clauses are sentences with subjects and verbs but without a complete thought expressed (subordinate clause). A dependent, or subordinate, clause needs to join an independent clause (a grouping of words that include at least a subject and verb and that comprise a complete thought) to become part of a complete sentence. Sometimes, two or more independent clauses can be conjoined (or coordinated) to become one sentence by the addition of a coordinating conjunction such as "and", "but" or other conjunctions. An increase in C/T shows that a writer or speaker is using more complex language, thus demonstrating more inclusion of subordinating conjunctions or relative pronouns. Further, it is a reliable measure of complexity (Foster and Skehan, 1996) and has been adopted in numerous studies. Next are examples of two written task repetition studies that have used the C/T measure.

Azizzadeh and Dobakhti (2015) used C/T as one of the syntactic complexity measures in a written task repetition study with 40 high-intermediate EFL learners in Iran, aged 18-25. Over 14 weeks, two groups (repetition versus no repetition) completed narrative writing tasks based on wordless picture stories. Although we can infer that the lexical complexity measures were mechanically calculated in their study, it was not explicitly stated whether the syntactic complexity measures were also calculated by mechanical detection. Their results showed that the experimental group significantly outperformed the control group in improved C/T.

Another written task repetition study that used this measure is Jung's (2013) where C/T were manually counted. Jung conducted the study with eight Korean university ESL students to investigate whether CAF increases after repeating tasks. This study also examined the effect of feedback. The results for C/T showed that, overall, the students who wrote their essays on the same topic had slight increases in C/T but the students who wrote their essays on a different topic had slight decreases. In terms of feedback conditions, all groups showed

improvements in C/T, thus feedback conditions did not impact performances in terms of this measure.

2.3.2.1.3 Dependent clause per clause (DC/C)

DC/C is another measure used to measure subordination. DC/C is measured by establishing the mean number of dependent clauses per clause. An increase in the number of dependent clauses per clause would suggest that there is more subordination. Next are two EFL writing studies that have used the DC/C measure.

Teng and Huang (2021) conducted a study of 352 intermediate EFL students from four universities in China. This study investigated whether metacognitive instruction and collaborative writing resulted in better writing, using various CAF measures. The students were divided into four groups based on the type of guidance they would receive. Such guidance included (1) "metacognitive instruction in a collaborative-writing setting; (2) metacognitive instruction in an individual setting; (3) collaborative writing; and (4) individual writing" (Teng and Huang, 2021, p. 1). Then the learners individually wrote an argumentative essay about the advantages and disadvantages of starting a business while attending classes. Results showed that there were slight variations of this measure among the groups. However, the results did not show a statistically significant effect between the groups in terms of DC/C.

Lu (2011) analysed 3,554 ESL writing samples from the Written English Corpus of Chinese Learners. These were essays that college students in China each independently wrote. Lu's study investigated the relationship between syntactic complexity measures and language development. The purpose was to identify whether any differences in the measures reached statistical significance between the student proficiency levels. Automated detection was used for all measures in this study. Results from the one-way ANOVA showed no significant differences among the students from the various colleges in terms of DC/C.

2.3.2.1.4 Dependent clause per T-unit (DC/T)

DC/T is another measure used to measure subordination. DC/T is measured by establishing the mean number of dependent clauses per T-unit. Next are two EFL writing studies that have used this measure. Rokoszewska's (2022) study was conducted among 100 secondary school EFL learners in Poland, whose written performances were collected over a timespan of three years. Over 1900 texts were retrieved from the Written Developmental Corpus of Polish Learners (WEDCPL) for this study. It aimed to determine how the monthly growth rates of various CAF measures changed over time on a series of tests. DC/T, among various other measures, was used to measure complexity. The monthly comparison was always compared to the previous

month, and there were 21 tests. Results from this study showed that the monthly growth rates for syntactic complexity (all measures together) increased steadily on tests 2, 17 and 19 but there was a decrease on test 10 in comparison to test 9.

Barrot and Agdeppa's (2021) study, introduced above, also used this subordination measure. Results showed that DC/T reached statistical difference when comparing this measure between proficiency levels in that there were steadily higher averages as the proficiency levels progressed.

2.3.2.2 Lexical complexity measures

In terms of assessing **lexical** complexity, a number of measures have been put forward in the literature.

Table 2-2 shows frequently used measures, along with examples of studies that have used them.

Table 2-2. *Lexical complexity measures*

Construct	Measure	Examples of studies that have used this measure
Lexical complexity		
Lexical diversity	Webtool measures for mean score for range of different words within a text	
	MATTR50	Kyle et al., 2023; Yang & Zheng, 2024
	VocD	Bui et al., ; Yang & Zheng, 2024; DeBoer, 2014
	TTR (Type/token ratio)	Gass et al., 1999; Yang & Zheng, 2024 Gass et al., 1999; Zhang et al., 2022; Manchon et al, 2023; Teng & Huang, 2021
Lexical sophistication	Sophisticated words & lexical sophistication (number of sophisticated words & mean score for sophisticated lexical words in proportion to total lexical words)	Manchon et al, 2023; Gass et al., 1999
Lexical density	Number of lexical tokens per total number of tokens	Kyle et al., 2023

2.3.2.2.1 Lexical diversity

A number of studies measure lexical diversity using various webtools. Lexical diversity is measured by establishing the mean score for the range of different words within a text. The

following are two examples of studies that investigated lexical diversity as part of a CAF study on EFL learning. It is worth noting that TTR and more recently Voc-D have long been the most popular measures used to establish lexical diversity measure. However, recently MATTR has become identified as one of the most reliable measures for this construct (Kyle, 2023), as it has been shown to be stable even when the texts scrutinized under this measure are very short (Bulté et al., 2024; Zenker & Kyle, 2021) – as also demonstrated by the two studies described below.

Kyle et al.'s (2023) study investigated the reliability and validity of lexical diversity measures, namely TTR, Voc-D, MATTR, and MTLD. 1,281 transcriptions from oral proficiency interviews (OPI) by intermediate Japanese EFL learners (based on the ACTFL Speaking Test) were extracted from the National Institute of Information and Communications Technology Japanese Learner English (NICTJLE) corpus. The study looked at the extent to which lexical diversity measures were predictive of OPI scores and the measures' stability across OPI tasks. Additionally, it looked at "the relationship between lexical diversity indices and text length in OPIs" (Kyle et al., 2023, p.6). Results showed that the MATTR and MTLD measures were reliable and valid because they were not affected by text length, while TTR and Voc-D were not reliable measures across the varying lengths of texts. According to McCarthy and Jarvis (2010), the reason why MATTR and MTLD are such reliable measures is they "capture unique lexical information" (McCarthy & Jarvis, 2010, p. 381).

Yang and Zheng's (2024) study investigated the effectiveness and sensitivity of various lexical richness measures among four Chinese EFL learner levels. It is important to note that lexical diversity is one aspect of lexical richness. Laufer and Nation (1995) define richness as the "degree to which a writer is using a varied and large vocabulary" (p. 307). Vermeer (2004) stated that lexical diversity is a popularly used measure of richness. Software such as Lexical Complexity Analyser, MATTR, MSTTR, and Coh-Metrix was used to calculate the lexical richness measures. 180 essays written by the four levels of learners were collected from an English corpus. Results from a one-way ANOVA test showed that MATTR was the most effective lexical diversity measure, regardless of essay length, confirming Kyle's (2023) research on improved lexical diversity measures.

2.3.2.2.2 Lexical sophistication

Lexical sophistication is measured by establishing the mean score for sophisticated lexical words in proportion to total lexical words. **Sophisticated words** is a calculation of the number of sophisticated words necessary to provide a proportion such as the number of words based on

similar time limits or word counts. Such ratios are necessary to compare learners' uses of advanced vocabulary to gauge improvements in language production performances.

Various webtools such as Lextutor provide the number of sophisticated words that come from word frequency lists, for example, COCA, New General Service List, Cambridge Corpus, etc. (Cobb, 2002; Heatley et al., 2002). The following is an oral task repetition study that used two measures for lexical sophistication: the number of sophisticated words and the mean score for the ratio of sophisticated words in proportion to the total number of lexical words.

Gass et al. (1999) conducted a study on the effects of oral task repetition on linguistic output with 103 native English-speaking university students who were studying Spanish in the United States. They investigated whether task repetition leads to more lexical sophistication and whether accuracy and/or lexical sophistication will be present to some extent when the language is produced in new contexts. There were three groups: two repetition groups (same content versus different content at each of the four performances) and one control group (watched same video only two times). The videos did not contain audio, i.e. some videos were silent from the start while with other videos, the audio portion was removed without affecting the comprehension of the video content. The number of lexical words and lexical sophistication were mechanically detected. All groups' scores improved, though the same-content group's lexical sophistication scores improved at a much higher rate than those of the two other groups.

2.3.2.2.3 Lexical density

Lexical density is the "percentage of lexical words in the text (as opposed to function words)" (Gass et al., 1999, p. 561). Lexical words are words that carry meaning (e.g. adjectives, nouns, verbs, places, proper names, etc.) and that also carry independent meanings. Function words, on the other hand, carry little lexical meaning except to show relationships between words. Examples are prepositions, articles, conjunctions, etc. Lexical density is a ratio of total lexical words (i.e., meaningful content words) to total words in a text. Language production with high lexical density thus means there is a large number of lexical words. Formal (or academic) speech or texts frequently carry higher lexical density than do informal/conversational texts (e.g. Maamuujav, 2021). Speech or text with low lexical density demonstrates that most of the words do not carry independent meanings.

Webtools such as Lextutor and Synlex Software have been used in various studies to calculate lexical density (Manchon et al., 2023; Gass et al., 1999). There is, however, some

criticism about measuring lexical density. Because lexical density is dependent on a text's syntax and cohesiveness, this measure does not always measure lexis. According to Laufer and Nation (1995), "[f]ewer function words in a composition may reflect more subordinate clauses, participial phrases and ellipsis, all of which are not lexical but structural characteristics of a composition" (p. 309). This exemplifies the importance for researchers to use other complexity measures to capture subordination.

Lu's (2012) ESL oral narrative study explored the relationship between lexical richness and language production. Four hundred eight oral narratives from the Spoken English Corpus of Chinese Learners were analysed. A webtool was used to measure lexical density as well as lexical sophistication and lexical diversity. The findings suggested a significant association between lexical richness and higher-quality oral performances.

Measuring Accuracy

Grammatical accuracy measures are identified as errors by type (e.g. subject-verb agreement) and by global errors (e.g. errors per 100 words). "Accuracy in writing has been captured by counting errors, calculating ratios of phrases, clauses, or T-units with and without errors, using holistic/analytic rating, and weighted error ratios" (Plakans et al., 2016, p. 164). Pallotti (2019) suggests evaluators use global measures to measure accuracy, notably errors per T-unit or per error-free T-unit. Also, as the total number of words that learners produce in their task performances is likely to differ between tasks, learners, etc., a measure of errors per 100 words has been used to control for those differences in text length. It is worth noting that in addition to grammatical accuracy, lexical is another aspect of accuracy. However, most research has looked into grammatical accuracy, and that will also be the focus of my own study.

In terms of assessing **accuracy**, a number of measures have been put forward in the literature. Table 2-3 shows key measures, along with examples of studies that have used them.

Table 2-3. *Accuracy measures*

Construct	Measure	Examples of studies that have used this measure
Grammar: • Subject-verb agreement • Verb tense • Verb form usage • Prepositions • Articles	Errors per 100 words (mean # of errors per 100 words)	Davison, 2024 ; Ahmadian & Tavakoli, 2010; Bui et al., 2019; Hassanzadeh et al., 2023; Jung, 2013; Muhammadpour et al., 2023;Bradley et al., 2018; Manchon et al., 2023; Skehan et al., 2023; Sanchez, 2023
	Errors per T-unit (E/T) (mean # of errors per T-unit)	Ahmadian & Tavakoli, 2010; Teng & Huang, 2021; Sang and Zou, 2023 ; Indrarathne, 2013
	Error-free T-units per T-unit (EFT/T) (mean # of error-free T-units per T-unit)	Sang & Zou, 2023; Ahmadian & Tavakoli, 2010; Jung 2013; Azizzadeh & Dobakhti, 2015; Indrarathne, 2013; Ghahderijani, 2021; Kizil, 2023; Sample & Michel, 2014; Teng & Huang, 2021

	Error-free clauses per clause (EFC/C) (mean # of error-free clauses per clause)	Jung 2013; Azizzadeh & Dobakhti, 2015; Kizil, 2023

Some examples of types of grammatical errors that have been measured in previous studies are subject-verb agreement (Ahmadian & Tavakoli, 2010; Sample & Michel, 2014; Kizil, 2023), verb tense (Davison, 2024; Kim & Li, 2024; Ahmadian & Tavakoli, 2010; Kizil, 2023), verb form (Muhammadpour et al., 2023; Ahmadian & Tavakoli, 2010), prepositions (Kizil, 2023; Davison, 2024, Kim & Li, 2024), and articles (Sample & Michel, 2014; Kim & Li, 2024; Larsson & Kaatari, 2020). Subject-verb agreement, verb tense, and verb form error types share in common that they are morphological. According to Bardovi-Harlig and Bofman (1989), learners of western languages tend to struggle with morphological forms even at high-intermediate or advanced levels whereas they tend to learn to use syntactic forms, for example, preposition usage, accurately much sooner, though this depends on the learners' L1 in terms of distance with the L2.Prepositions and article error types share a commonality in that they both help make sentences more understandable. However, the similarity between prepositions and articles stops there – articles describe nouns while prepositions show relationships between objects. According to Miller (2015), because errors in article usage do not prevent communication though sometimes can cause miscommunication, potentially, L2 learners whose L1 does not have articles might neglect making the effort to learn correct article usage. However, Master (1997) stated "imperfect control [of the use of articles] may . . . suggest imperfect knowledge", leading to the perception that the learner does not have an adequate grasp of their subject (p. 216).

VanPatten (2012) explains how learners tend to focus primarily on meaning rather than on form. As a result, morphological forms tend to be frequently ignored by foreign language learners, e.g. "while morphemes and inflections might be perceived and/or noticed, they are not processed" (p. 270). Meanwhile, Ferris (1999) suggests that errors with verb tense and form are treatable since they occur in a "patterned, rule-governed way" (1999, p. 6), and research on oral CAF as well as studies on written CF lend support to this claim (e.g. Bitchener et al., 2005; Yang & Lyster, 2010).

It is worth noting that according to Ellis and Barkhuizen (2005), learners will not improve in all linguistic features at the same rate: "learners do not acquire grammatical features concurrently. Rather, some features are acquired early and others late" (p. 59). To that

end, many studies provide the total number of errors rather than by error type. In this case, "global measures, by contrast, examine the text or transcript in its entirety…the segmentation may be made by dividing the data into units, [for example], per 100 words, per T-unit, per clause, etc…." (Foster & Wigglesworth, 2016, p. 102).

Next are some examples of studies that have used global measures.

2.3.2.2.4 Errors per 100 words

Errors per 100 words is measured by establishing the mean number of errors (all error types taken together) per 100 words. The following is a written task repetition study that used this measure.

Davison (2024) conducted an EFL writing study on collaborative writing with 128 upper-intermediate EAP students during a two-semester period in the United Arab Emirates, aged 19-21, whose first language was Arabic. The study investigated how writing changes in terms of CAF after completing collaborative versus independent writing activities. During the two semesters, four groups of students completed collaborative writing activities and four other groups wrote independently. The students in the second-semester groups were different from those in the first semester. All groups were taught by the same teacher.

In terms of the types of errors studied in Davison's (2024) study, the following were manually counted with a second rater: verb form and tense, subject-verb-agreement, articles, prepositions, word order, pronouns, and comparative and superlative adjectives. In terms of results for the mean number of grammatical errors/100 words, the findings in the post-test (compared to the pre-test) showed statistically significant effects of time, i.e., a significant decrease in the mean number of grammatical errors per 100 words for all groups taken together over time. In terms of comparing the two types of interventions, i.e., collaborative and independent writing groups, there were no statistically significant findings in the differences in grammatical errors per 100 words; there was no interaction effect between time and treatment.

2.3.2.2.5 Errors per T-unit (E/T)

Errors per T-unit is measured by establishing the mean number of errors per T-unit. The following is an integrated reading-to-write study that used this measure. Sang and Zou (2023) investigated the impact that teacher-scaffolded feedback and collaborated dialogue has on EFL reading-to-write in terms of complexity and accuracy of the written output. The participants were fifty-nine 18-year-old first-year lower-intermediate to intermediate EFL university students in China who had been learning English for 10-12 years. They were split into two groups; both groups completed the same untimed reading-to-write tasks: the experimental

group received teacher-scaffolded feedback while the control group received only conventional feedback. The experimental group received real-time interactive joint production feedback, i.e. integrated group work: peer feedback, then immediate detailed teacher feedback, i.e., scaffolded, during the post-reading tasks. The control group received only written feedback from the teacher. This study was conducted over an 18-week period. Two researchers manually counted and coded all grammatical and lexical errors. The results revealed that both groups made significant improvements by making fewer E/T, although the experimental group improved to a statistically significant larger extent than the control group did, i.e. there was an interaction effect of time and treatment in terms of E/T..

2.3.2.2.6 Error-free T-units per T-unit (EFT/T) & Error-free clauses per clause (EFC/C)

Error-free T-units per T-unit is measured by establishing the mean number of error-free T-units per T-unit. Error-free clauses per clause is measured by establishing the mean number of error-free clauses per clause.

The following are examples of EFL studies that used both of these measures. Ghahderijani's (2021) study investigated the impact that different types of feedback made in terms of CAF among 30 intermediate EFL learners in a language school in Iran. The learners were female, aged 16-19 years old. They were split into two groups that completed the same writing task over eight sessions: the experimental group received detailed feedback and collaboration from the teacher and from peers whereas the control group received only brief corrective feedback. The types of accuracy errors being under scrutiny were word forms and verb tenses, which were manually counted, however combined into one measure of accuracy errors and not kept separate. The results revealed that in terms of accuracy (EFT/T and EFC/C), there was a highly significant difference in mean scores between the two groups, with the experimental group having the higher mean than the control group.

Azizzadeh and Dobakhti's (2015) study, introduced earlier, also used EFT/T and EFC/C as measures in the accuracy component of their research. All types of errors were manually counted (with a second rater) except for the following types of errors as long as they did not affect comprehension: capitalization, prepositions, punctuation and lexical words, i.e. these aforementioned types of errors were not counted at all if the errors did not impede comprehensibility. It was found that even though the experimental group had somewhat higher average means of EFT/T and EFC/C than the control group did, the differences were not statistically significant.

CAF researchers can choose between using E/T versus EFT/t based on their focus. E/T measures the overall density of errors such as average number of errors per chunk. EFT/T focuses on whether learners produce correct units, which would be applicable for higher-level learners. Also, researchers might select measures that align with earlier studies that they replicate.

Measuring Fluency

Fluency is typically measured by length of production and errors, based on the premise that learners with high proficiency will produce more written or spoken words with fewer errors. Thus, fluency measures are typically based on the amount of L2 speech or writing that students can provide within a specified amount of time (Yoon & Polio, 2019). For example, Fellner and Apple (2006) describe fluency as measured by "the number of words produced in a specified time frame, together with lexical frequency, irrespective of spelling and content, provided that the writer's meaning is readily understandable" (p. 19). In addition, some fluency measures capture the learner's ability to produce elaborate t-units and error-free language (Wolfe-Quintero et al., 1998 – for the context of written production).

Fluency ratios are typically calculated as words per T-unit (W/T) and words per error-free T-unit (W/EFT). In order to calculate these fluency ratios, researchers need to first calculate words per text and T-units per text. Nevertheless, fluency ratios are seen as effective because they measure the student's ability to produce longer and more advanced sentences with improved fluency in writing or speaking. Many studies confirmed T-units as a reliable measure of syntactic development because T-units are easy to identify and measure (Wolfe-Quintero, Inagaki and Kim, 1998; Gass & Selinker, 2001; Mackey & Gass, 2005). ~~I~~n addition, it is a reliable indicator of syntactical development (Larsen-Freeman, 1983; Bardovi-Harlig & Bofman, 1989), and integrating the T-unit measure of syntactical development helps confirm a learner's proficiency in terms of accuracy and/or fluency.

In terms of assessing **fluency**, Table 2-4 shows key measures found in the literature, along with examples of studies that have used them.

Table 2-4. *Fluency measures*

Construct	Measure	Examples of studies that have used this measure
Written language proficiency	Words per T-unit (W/T) (mean # of words per T-unit)	Ghahderijani, 2021; Phuoc, 2022; Rathi; Tai, 2015; Hattingh, 2007; Indrarathne, 2013; Roko, 2022; Zhang et al., 2022
	Words per error-free T-unit (W/EFT) (mean # of	Hattingh, 2007; Jiang, 2013; Meletiadou, 2021; Ping-Ju, 2020; Stell, 2018

words per error-free T-unit)	Barrot and Agdeppa, 2021; Kizil, 2023; Teng & Huang, 2021; Ghahderijani, 2021
Words per text	Barrot and Agdeppa, 2021; Kizil, 2023; Teng & Huang, 2021
T-units per text	

2.3.2.2.7 Words per T-unit (W/T), words per text, and t-units per text

Words per T-unit is measured by establishing the mean number of words per T-unit. Following Zabihi (2018) and Larsen-Freeman (2009), L2 writing fluency is measured in many studies by the average number of words per T-unit.

Two recent EFL writing studies, Ghahderijani (2021) and Barrot and Agdeppa (2021), introduced earlier, used W/T as one of the fluency measures in their research. Results from Ghahderijani's study revealed that in terms of fluency, there was a statistically significant difference in mean scores between the two groups, with the experimental group having a higher mean than the control group. Results from Barrot and Agdeppa's study revealed that in terms of fluency (words per text and t-units per text), the mean for words per text reached statistical differences when comparing this measure between proficiency levels in that there were steadily higher averages of words per text as the four proficiency levels progressed. There was a statistically significant decrease in the mean for T-units per text across the proficiency levels, i.e., fewer T-units per text.

2.3.2.2.8 Words per Error-Free T-unit (W/EFT)

Words per error-free T-units per T-unit is measured by dividing the total number of words (contained only within the error-free T-units) by the number of error-free T-units. It is a popularly used measure that accounts for the number of words, T-units, and errors.. It is necessary to consider that when a learner produces more words, it is not automatic that accuracy improves concurrently. Because T-unit count measures alone do not suffice for syntactic development, error-free T-unit analysis is an additional measure used to assess the number of errors in relation to the sentence length. Wolfe-Quintero et al. (1998) indicate error-free t-units help identify fluency because they "capture the fluency of a writer within the context of writing accurate sentences" (p. 15). Davison (2021) stated: "Words per error-free t-unit includes elements of accuracy and complexity, but it also highlights the writer's ability to write longer, more elaborated sentences that are error-free within a given period of time" (p.84). This helps confirm whether length increased with improved linguistic accuracy. In this

case, only the T-units that do not contain errors can be included in the equation. Next is an example of an EFL study that used this measure.

Davison (2024), as introduced in section 2.3.2.2.4, conducted an EFL writing study on collaborative writing with 128 upper-intermediate EAP students during a two-semester period in the United Arab Emirates, W/EFT was one of the fluency measures used in this study. The number of T-units were manually counted with a second rater. The number of words were mechanically counted. In terms of results for the mean W/EFT, the findings showed statistically significant effects of time, i.e., an increase in the mean W/EFT for the groups taken together over time. In terms of comparing the groups, there was no statistically significant findings in the difference in the fluency of their writing. In addition, there was no interaction effect between time and treatment.

2.3.3 Task Repetition as a pedagogic technique

Now I explain the rationale underlying the use of task repetition as a pedagogic technique in more detail, and its theorized effect on CAF.

Bygate (1996) argued in the context of oral task repetition that if learners are given the opportunity to repeat a task, they will gain oral accuracy because they are familiar with the task content by having completed the first task completion, thereby allowing them to shift their focus increasingly the next time on producing the correct L2 formation. In this case, they shift their attention to selecting the language, thereby monitoring its complexity, accuracy and fluency (Bygate, 1999). Bygate (2001) sums up: "part of the work of conceptualization, formation and articulation carried out on the first occasion is kept in the learners' memory store and can be reused on the second occasion" (p. 29). This suggests that task repetition facilitates learners' use of the rule-based system, thus helping improve CAF performances.

Two highly competing hypotheses that are highly relevant to task repetition and that identify ways that complexity may impact attentional allocation during task performance are the Trade-Off Hypothesis and the Cognition Hypothesis. As explained earlier in Section 2.1, the Trade-Off Hypothesis (Skehan, 2009, 1998a), proposes that "due to capacity limitations, speakers must divide their attentional resources between all the processes a task requires…If various task demands exceed the available resources, [then complexity, accuracy and fluency compete] with each other" (Sample & Michel, 2014, p. 27).

Robinson's (2007) Cognition Hypothesis, however, is a theory that competes with the Trade-Off Hypothesis. The Cognition Hypothesis claims that pedagogic tasks should be sequenced for learners on the basis of increases in cognitive complexity and that the increase

of a task's cognitive demands will push language learners to make significant improvements on accuracy and complexity, thus encouraging procedural repetition as another type of pedagogic task repetition technique. This theory hypothesises no competition between complexity and accuracy because these areas of L2 production are closely linked. Further, Robinson predicted that more complex tasks would result in greater accuracy and lexical complexity but less grammatical complexity (Robinson, 2005). Consequently, it should entice learners to increase focus on memory of L2 input for longer-term retention of the input, and it claims that performing of complex sentences will lead to automaticity of the complex task output.

These two theoretical models which have initially been put forward to model task effects in speaking have also been adapted to conceptualize task complexity research in writing (Kuiken & Vedder, 2008). To extend the point made about memory and planning, Kellogg's 1996 model of writing stipulated that memory aids in planning ideas for written sentences, thereby supporting that students retrieve topic knowledge from memory and then use it to complete a complex cognitive task (Ransdell et al., 2001; Kellogg et al., 2016). Essay writing requires students to use many cognitive processes including planning what to write, developing sentences, and revising what they write (Hayes, 1996; Kellogg, 1996; Levy & Ransdell, 1995). These cognitive processes in this writing theory give rise to my hypothesis that students improve in their writing performances as a result of repeating the same task, but with their ability to shift their focus on the process itself when repeating the task rather than on learning what the task is about at the first writing performance.

According to Revesz et al. (2017), Kellogg's model, originally developed to account for first language (L1) writing, lends itself well to studying L2 writing processes, notably as it makes detailed predictions about linguistic encoding processes, which, in comparison to L1 writing, "are likely to generate considerable cognitive demands for L2 writers" (Révész at al, 2017, p. 5). Though Kellogg's model, similar to other writing models, makes no direct predictions regarding the relationship between specific task manipulations and L2 writing processes and outcomes, these correlations have received considerable theoretical and empirical attention in the area of L2 speech production (Skehan, 2014; Robinson, 2011).

While the above theories hang on oral repetition, the focus of the effects of the task repetition stages from these studies frame my study to assess students' written language production. Further, it is important that written task repetition research is conducted to broaden as well as verify the theories. Having introduced the main rationale for task repetition, and

definitions of the CAF concepts, I will now review in more depth the literature on oral task repetition and CAF, and then written task repetition and CAF.

2.3.4 Oral task repetition and CAF

Previous task repetition studies have shown that repetition of the same task has different effects on learners' repeated oral production. Task repetition has been found, overall, to be more effective in promoting improved CAF performances in oral language output (e.g. Ahmadian & Tavakoli, 2011; Bygate 2001; Sample & Michel, 2014). There are many studies on oral task repetition and CAF (e.g. Fukuta, 2016; Zhang et. al, 2023; Nguyen et. al, 2023). Below, I will review a selection in detail. This selection of studies below provides various examples of methodologies used such as same- versus procedural repetition, number of repetitions, types of intervention for the experimental groups, etc. In addition, the selection of studies exemplifies variations in findings. For example, results from some studies show significant differences between groups or methodology while results from other studies suggest no significant differences. Also, findings from many studies show variations in the specific CAF measures where there were significant differences. In addition, where there were some trade-off effects in terms of CAF, complexity was generally stronger than accuracy and fluency, but that was not always the case for each finding.

Bygate (1999) advocated that oral task repetition has many beneficial effects on L2 performance. Building on Levelt's (1989, 1999) speech production model, Bygate (2001) argued that when L2 students complete a task the first time, their speech production system needs to use all the relevant language processing steps while under time constraints. Based on this limited attentional model of speech production (Kormos, 2006; Skehan, 2009), it could be argued that on the first (or only) performance of a given task, learners have to strategically use their attentional resources to conceptualize, formulate and monitor their L2 production. Similarly, students also have to overcome challenges of handling the breakdown of their L2 performance due to incomplete lexical or syntactic proficiency (Dörnyei & Kormos, 1998).

When learners repeat tasks, they already become familiar with the content the first time when they do most of their conceptualizing, formulating, and articulating of language, thus enabling them to redirect their attention from content to proper language usage when repeating the task, leading to more advanced complexity, fluency and accuracy as hypothesised by Bygate (1999) and Bygate and Samuda (2005). There is, however, an exception to the repetition effects: Bygate and Samuda (2005) suggest that such an effect may be minimal if student performance for the same task is automated, i.e., it might be easy for the learners in the

first place or if they do not put in effort to improve, thereby different from the typical repetition effect. In this case, learners are also merely reusing information they had released from memory when producing language.

Results from empirical research that compares and contrasts L2 output after timed intervals between initial task completion and task repetition help confirm the effects. Studies show that task repetition improves CAF performances even though each measure improves sometimes at the expense of another. Bygate (1996) was among the first to empirically research oral task repetition and its effect on L2 cognitive processing. Bygate (1996) asked an L2 learner to watch the same cartoon and then immediately speak about what occurred. This task was repeated once three days later. The conditions were identical on each occasion, i.e., the task, the task instructions, the interlocutor, and the room. The only difference was the learner's familiarity with the story on the second occasion. Bygate found that this task repetition led to improvements in the learner's oral fluency and accuracy. Results also showed repetition of several phrases from the first narration, which likely is L2 production resulting from discourse planning between tasks or memory from the first narration. A major limitation of this first study was, however, that it involved only one learner.

A couple of years later, Gass et al. (1999) conducted a study on the effects of task repetition on linguistic output using a much larger sample – 103 English-L1 university students who were studying Spanish. Like in Bygate's (1996) study, the participants were asked to narrate a short audio-free video by telling what was taking place. There were three groups to which the students were randomly assigned: two experimental groups and one control group. The two experimental groups completed four performances within two weeks, with two-to-three-day intervals between the first and third performances, then one week between the third and fourth performance. One of the experimental groups (same content group) watched the same video the first three times, then a new video at the fourth performance. The other experimental group (different content group) watched the same video as the same content group the first time, then a different video the remaining times, so a different video at each of the four time points. At the fourth performance, both experimental groups watched the same video. The control group watched a different video at two different times. The first time the control group watched a video was at the same time that the two experimental groups watched a video for the first time (all three groups watched the same video at the first video viewing). The second time the control group watched a video was at the same time that the experimental groups watched a fourth video (all three groups watched the same video at the fourth viewing, but it was the second time that the control group watched a video because that group did not

view a video during the second and third times that the experimental groups watched a video. The video that the control group watched at the second viewing was a different video than they watched at their first viewing. However, the second video that the control group watched was the same video as the one that the experimental groups watched on their respective fourth viewing. So, the three groups watched the same video at the first performance and the same at the last performance, but the video shown at the last performance was different than any of the videos that the groups watched during earlier performances.

Results from Gass et al.'s study showed that the same-content experimental group showed a higher increase of fluency than did the different-content experimental group between Time1 and Time3, as well as between Time1 and Time4. The same-content group's fluency scores (a holistic measure of magnitude estimation for improvement in overall proficiency compared to Time1 performances) at Time4 were higher than that of the different-content group. The control group's fluency improved only slightly between their two performances. Regarding accuracy, all of the groups improved from Time1 to Time3, and from Time1 to Time4. However, there were two measures with different results, the use of the Spanish verb 'to be', "ser" and "estar." For "ser," from Time1 to Time3, the scores remained about the same, and from Time1 to Time4, there were slight improvements for all groups, with the same-content group's improvement percent higher than that of the different-content group. For "estar," from Time1 to Time4, all groups improved, with the different-content group's improvement percent higher than that of the same-content group. In addition, the control group's improvement percent was very similar to that of the same-content group. Regarding complexity, only the Time1 to Time4 groups' performances were measured because it was at Time4, not Time3, that all groups watched the same video as one another.

All groups' scores improved, though the same-content group's lexical sophistication scores improved at a much higher rate than those of the two other groups. In sum, repeating a task had an effect on the students' overall oral proficiency and lexical complexity, but a partial effect on accuracy (with "estar" but not "ser"). Taken together from start to finish, results showed a systematic increase in CAF. Because accuracy had only a partial effect, this suggests the possibility of a trade-off effect.

In 2001, Bygate built on and expanded his initial 1996 research on oral task repetition by conducting a study where he compared 48 university ESL learners' oral performances on two types of tasks: a narrative and an interview. This study researched the effects of task-type practice and task repetition on repeated oral performances. There were three groups in which the students were randomly assigned: two experimental groups (narrative group and interview

group) that repeated the task four times (five performances) over a 10-week time period and one control group that completed the task only two times (beginning and end of the 10-week period). None of the groups received feedback between performances. The narrative experimental group was asked to talk about what was going on during dialogue-free cartoons; the interview experimental group was asked to talk about the way their lives were similar or different from photos that depicted various aspects of life. All groups completed both a narrative and an interview task at Time1 and Time5. Then during 2-week intervals, the two experimental groups each completed two tasks but of the same task-type, one that was a repetition of their task at Time1 and one that was different. At Time5, all groups completed two narrative and two interview tasks (4 tasks total at Time5), with two of those tasks being the same as the tasks at Time1.

Bygate (2001)'s ANOVA results showed that repeated task-type practice and task repetition affected CAF measures on students' oral performances in that complexity and fluency improved but at the expense of accuracy. This finding suggests a trade-off effect. There were statistically significant findings for complexity for both task types (narrative and interview) but no statistically significant findings for accuracy. Results also showed that "the fluency correlations suggest considerable consistency of performance, while correlations on the complexity and accuracy measures suggest that consistency is strongest across tasks of the same type or on familiar tasks" (Bygate, 2001, p. 36). In terms of fluency, there was a statistically significant finding showing less fluency and greater complexity with the interview task type than with the narrative task type. Given that the results showed a strong effect from task repetition with a weaker effect from task-type repetition, these findings suggest that task repetition can help ESL learners develop their language proficiency despite that some of the CAF measures compete with one another in the repeated performances.

Ahmadian and Tavakoli (2011) continued the work on the narrative task type, but further refined it by exploring the combined effect of task repetition and planning conditions. They examined the CAF effects of simultaneous use of task repetition and careful versus pressured online planning on L2 learners. Sixty Iranian intermediate-level adult female EFL learners, who had been studying EFL for six months, were randomly assigned to one of four groups based on task planning and repetition conditions (careful online planning with repetition, pressured online planning with repetition, careful online planning without repetition, and pressured online planning without repetition). Several pretests were conducted at the beginning so that the researchers could identify initial differences across the groups as well as to verify equivalency. The result of an accuracy pretest showed no statistically significant difference across the groups.

Similarly, a pretest for fluency was also conducted, and results showed that the participant groups overall were equivalent. Further, the listening subtest of TOEFL was conducted to determine the participants' online processing abilities, and its results showed that all groups were fairly equivalent in this area.

In Ahmadian and Tavakoli's (2011) study, participants performed an oral narrative task. First, they watched a 15-minute silent film, then within the various group conditions, they were asked to retell what took place in the film. The groups were informed before the film started that they would be asked to retell the story as if they were telling it to someone who wanted to know the details. The two repetition groups repeated the same task after one week, and participants in the two groups without repetition conditions performed the task only once. The careful planning group that was under the repetition condition had no time limit while the two pressured planning groups were given a 6-minute time limit to complete the task.

Results showed that simultaneous use of careful planning and task repetition positively affected CAF more so than for any other group. In addition, in terms of task repetition, one-way ANOVA results confirmed that there were statistically significant findings that showed that repeating this task enhanced the oral narrations in the performances from all repetition groups. For the repetition groups taken together, results from task repetition revealed performance increases in complexity and fluency while at the same time, there were no significant changes in accuracy.

In terms of careful versus pressured online planning conditions, the findings for complexity and accuracy were statistically significant for participants under the careful planning condition, i.e. the careful planning groups produced more complex and accurate narration than did those under the pressured planning condition. Regarding fluency, the findings from the fluency measures' means for the pressured planning groups were statistically significant, i.e. the pressured planning groups produced more meaningful syllables per minute than did the careful planning groups.

In terms of the effects of simultaneous use of task repetition and careful online planning, this combined usage resulted in enhanced performances in CAF compared to Time 1. For example, this group completed the narrations by using more complex language along with enhanced accuracy, i.e. they spoke using more correct verb forms and error-free clauses. Fluency performance for all groups was somewhat the same. Additionally, in comparison to the other groups, the effects from the combination of careful planning and task repetition for this group were higher on complexity than if the other groups performed under either of the task repetition or careful planning on its own. Unlike the results from Bygate's (2021) study,

the findings from the task repetition groups show that there were no trade-off effects in the CAF measures.

Next is a study that compared the effects of task repetition and procedural repetition on language learners' L2 oral production. Kim and Tracy-Ventura's (2013) study compared 36 Korean junior high school EFL learners who were split into two groups: (1) task repetition and (2) procedural repetition groups. All participants completed a pretest, three collaborative tasks, and two post-tests over a four-week period. The task repetition group repeated the same information-exchange task procedure with the same content three times while the procedural repetition group repeated the same information-exchange task procedure with different content. The oral production on the pre- and post-tests were analysed using CAF measures. The tests entailed pictures that showed various scenarios such as hosting an American friend, planning a school event and discussing mayoral candidates. The learners, in pairs, then discussed the content, then generated information reports in pairs.

Results revealed, regarding complexity, that both methods of repetition were beneficial for improving linguistic forms and that the procedural repetition group additionally improved in syntactic complexity. However, the results showed that repetitions did not impact improvement of lexical complexity for either group. Regarding accuracy, overall, both groups did not show any significant increase in their global accuracy with the exception of simple past tenses where both groups improved significantly in their correct use of this verb tense, thus suggesting no group difference for accuracy apart from this specific measure. Regarding fluency, there was no significant increase in speech rate between the two repetition treatments, i.e., no time or group effect..

Sample and Michel (2014) continue the work on oral task repetition by investigating the effects on language output, on a different task type, with two repetitions (so three performances). They conducted an investigative study on the interactions among CAF measures in repeated oral task performances of spot-the-difference tasks by six 9-year-old language learners (Chinese-L1) in an after-school English course and whose EFL oral proficiency levels were advanced beginner to lower-intermediate. An original picture and three slightly different pictures were used for the task. The learners were grouped into three pairs, then each pair was asked to identify six differences that they spotted between the two pictures they were given. There were different pairings of learners at each performance, and, at repeated performances, there was one picture that the pairs saw during the previous performance. These were two-way (pair) tasks, and both learners in each pair were encouraged

to participate equally. After each task completion, the pairs compared each other's pictures. The first performance and the two repetitions took place over a three-week period.

Sample and Michel's (2014) results revealed a trade-off effect from Time1 to Time2. Then like earlier studies (Bygate, 2009; Hawkes, 2009, 2012), the trade-off effect disappeared at Time3. Namely, from Time1 to Time2, learners who used complex structures made more grammatical errors, i.e. a trade-off between structural complexity and accuracy. Also, learners who used more elaborate vocabulary also made more pauses during their narrations, i.e. a trade-off between complexity and fluency. However, at Time3, the trade-off effects did not appear. This finding suggests that the more times that learners repeat a task, the less competition between the CAF measures in terms of improvements in language production. This finding also suggests that once learners become familiar with the tasks, they can focus their attention on all CAF dimensions at the same time, thereby corroborating the Trade-Off Hypothesis at the second performance, and backing earlier research in support of further repetitions to minimize a trade-off.

More recently, Bui et al. (2019) investigated the effect of spacing in oral task repetition (i.e. time intervals between repetitions). They examined the CAF effect of task repetition on 71 first- and second-year college students in an EFL course in Hong Kong under five task performance conditions with different time intervals, i.e. immediate, one-day, three-day, one-week, and two-week intervals. The students' L1 was Chinese, and they had not spent more than three months in an English-speaking country. The students were randomly divided into five task repetition groups distinguished by different time interval periods between repetitions. This was a picture description task where each group was asked to speak about the picture, that depicted a household setting. Each group then repeated the task at their assigned task repetition time.

Bui et al.'s results suggested that regardless of the timing variable, task repetition had an overall positive effect on L2 learners' CAF measures. First, in terms of complexity, there was higher structural complexity but there were no statistically significant findings between the time interval spacings. However, repetition did not change lexical diversity. In regard to longer space intervals, the largest gain occurred with the one-week interval group who produced longer-length AS units when repeating the task. Although the findings showed no interaction effects between spacing and task repetition for complexity, task repetition raised structural complexity. Second, in terms of accuracy, there was no effect between task repetition and spacing. Third, in terms of fluency, "both task repetition and spacing significantly affected speech rate" (p. 7). Taken together, all groups had a higher speech rate during the repeated

performance. Regarding time intervals, the greatest improvement in fluency occurred with the group that had the three-day interval. The findings showed that the speech rate between the one- and two-week intervals were similar.

Next is a sequence of three studies exploring the potential added benefit of transcriptions to task repetition effects.

In Mennim's (2003) study, three first-year upper level EFL students at a Japanese university transcribed five minutes of the first performance of a tape-recorded oral presentation that they did. They then received the teacher's feedback based on the transcriptions one week later. One week after they received feedback on their transcripts, they were asked to repeat the oral presentation. This study investigated whether students benefitted from rehearsing the task between the two performances. Mennim was looking at whether this rehearsal, along with feedback on the short transcriptions, would help students improve on language form as a result of a shift of attentional capacity as seen with task repetition. Results showed that on the repeated performance, the students used more correct linguistic forms and reformulations, i.e. improvements in pronunciation, grammar, and content organization. A major limitation of this earlier study was, however, that it involved only one group (so no control group) and the number of participants was minimal.

Hsu (2019) conducted a task repetition study by comparing three groups' oral performances based on CAF measures using three sets of picture-based narrative tasks, with six related pictures for each task, over a three-week period. This study not only looked into the role of learner transcriptions (where one group was asked to transcribe what they heard based on audio recordings of their narrations where they discussed the pictures), but also the difference between exact prompt repetition versus new prompt (task-type repetition only). Thirty-nine Taiwanese university EFL students were placed randomly into three groups. The task repetition post-transcription group repeated tasks 1 and 2 and their post-transcriptions of their oral performances one time, one week apart. The task repetition group (no post-transcription) repeated tasks 1 and 2 one time, one week apart. The control group completed three different tasks, one week apart from one another. All groups were given unlimited time to complete the narrative tasks. At Time1, all groups completed narrative task 1, and the post-transcription group was given one week to transcribe their narration recordings and to create separate transcriptions that reflected the corrections of errors they believed they made on the transcripts. There was a one-week time gap between repetitions. At Time2, all groups completed narrative task 2; the two repetition groups repeated narrative task 1, and the post-transcription group further completed the two transcriptions based on their own oral

performance of narrative task 2. At Time3, all groups completed narrative task 3; the repetition groups also repeated narrative task 2.

Results from an ANOVA test showed no statistical significance between the three groups' performances in Hsu's study in regard to CAF measures at Time1, therefore the three groups' levels of oral performance were similar at that point. In terms of repeated task condition, the results revealed that for complexity, there was no significant difference between the two repetition groups, thus post-transcribing did not result in more complexity in L2 production. For accuracy, there were significant differences between the two repetition groups, i.e. the group that transcribed spoke with more accuracy when repeating the task than did the repetition group that did not, thus suggesting that post-transcribing helps improve accuracy. In terms of CAF in the new tasks, the results revealed that for complexity and fluency, there were no statistically significant differences between the groups. As for accuracy, regarding error-free clauses, there were statistically significant differences between the three groups, i.e. the repetition group that transcribed had more error-free clauses than did the repetition group that did not transcribe, and both repetition groups made more error-free clauses than did the control group. Regarding verb form usage, there was no statistical significance between the three groups, but there was statistical significance for the two repetition groups, i.e. the repetition group that transcribed made fewer verb form errors than did the repetition group that did not transcribe. However, there were no significantly different findings between the repetition groups and the control group. In sum, results showed that the group that combined task repetition and transcribing was more effective in helping learners maintain their accuracy gains that carried over to the new tasks. Further, like several of the earlier studies discussed in this section, these findings where there were significant improvements for specific accuracy measures at the repeated performances at the expense of complexity suggest a trade-off effect at the second performance. Similar gains to Hsu (2019) in oral performance through post-transcriptions align with Mennim's (2003) much earlier task repetition study.

Like Mennim (2003) and Hsu (2019), Hassanzadeh-Taleshi et al. (2023) also investigated the effects in terms of CAF of oral task repetition and post-transcription of learners' completed narrations. Hassanzadeh-Taleshi et al.'s study combined task repetition with immediate post-task transcribing while also varying immediate and delayed oral repetitions of the same task (based on a silent cartoon without dialogue). Thirty-eight intermediate Iranian university Teaching English as a Foreign Language (TEFL) majors were placed randomly into two groups: one group that repeated the task and one group that transcribed and then repeated the task. Results from an independent t-test showed no statistical

significance between the scores from a placement test taken before the task performances began, thus both groups had similar oral proficiencies at the start.

In Hassanzadeh-Taleshi et al.'s study, for the first repetition, the task repetition group repeated the task immediately after their initial performance; the task repetition and transcription group transcribed their initial performance in the same sitting on the same day before immediately repeating the task. Completion time for both groups was unlimited, but the cartoon was not replayed for the immediate repetition. For the second repetition, one week later, all participants repeated the same task. Results for both immediate and delayed task repetition revealed that there were no statistically significant differences between the performances in terms of CAF measures for either group.

This next oral task repetition study presented here looked at the impact of a cognitive factor on the participants (working memory). Muhammadpour et al.'s (2023) study examined the effects of task repetition on CAF at three different time intervals. Thirty-six intermediate Iranian male university EFL students were split evenly into two groups based on their working measure capacity scores from a Speaking Span Test, i.e. a low working memory capacity group (LWM) and a high working memory capacity group (HWM). Unlike several other task repetition studies discussed above, there was different spacing for the time intervals between the first and the repeated performances. At Time1, both groups watched a silent three-minute cartoon, and then performed a narrative task by retelling what took place in the show. There was no time limit for the narration. Then they immediately repeated the same task. Like Hassanzadeh-Taleshi et al.'s study, the cartoon was not replayed. Three days later, both groups repeated the same task without viewing the cartoon. Then one week after their first and immediate repeated performances (and so four days after their previous repetition), both WM groups repeated the same task again, also without viewing the cartoon.

The MANOVAs run by Muhammadpour et al. (2023) revealed that there were no statistically significant effects for CAF between the two WM groups at their first task performances. In terms of immediate task repetition, for complexity, there were no statistically significant differences; for accuracy, there was a statistically significant effect between the groups for correct verb forms and the LWM group performed higher than the HWM group, but there was no significant effect for error-free clauses or error-free AS units; for fluency, there was no significant effect between the two groups. In terms of repetition after three days, there was no significant effect between the two groups for CAF. In terms of repetition after one week, for complexity, there were significant effects for the amount of subordination and lexical diversity, and the LWM group performed higher than the HWM group; for accuracy, there was

a significant effect for error-free clauses, and LWM group performed higher than the HWM group; for fluency, there was no significant effect between the groups. It is worth noting that taken together across the repetitions, the MANOVA results showed a significant effect for lexical diversity, and LWM group performed higher than the HWM group. These findings suggest that (1) the opportunity for immediate task repetition helped the LWM group produce more correct verb forms more so than the HWM despite both groups being similar at the start, and (2) like the results from several other task repetition studies introduced earlier in this section, there were trade-off effects among the CAF measures.

Next is Nguyen et al.'s (2023) study with a much larger sample of an oral task repetition study of 27 second-year English majors (19 years old) at the B2-level Common European Framework of Reference for Languages (CEFR) scale at a university in Vietnam (randomly and evenly split into three groups) over a four-week period. This study investigated the impact that task repetition had in conjunction with post-task-teacher-corrected transcribing (feedback) on speaking performances in terms of CAF. The study looked at whether task repetition combined with post-task-teacher-corrected transcription helped improve language performances in terms of CAF and whether this same combination improves in terms of CAF when students perform new tasks. Each task entailed a narration, no time limits, that described a different photo that related directly to the same thematic topic each time, environmental problems. Students were asked to respond to an open-ended question about their judgement regarding environmental concerns depicted from the photos.

Group 1 consisted of students who repeated the task and then received post-task-teacher-corrected transcription; Group 2 consisted of students who repeated the task but did not receive post-task-teacher-corrected transcription; and Group 3 (control group) completed a new task each time and did not receive post-task-teacher-corrected transcription. Group 1 completed each task, and then immediately transcribed their recordings that they then sent to their teacher for corrective feedback. Groups 1 and 2, both repetition groups, repeated the tasks one week after the previous performances on the same task, and then immediately completed a new task after that.

During week 1, a diagnostic test was given to determine language proficiency levels. During week 2, group 1 completed Task 1, then immediately transcribed the narration recording and sent it to the teacher. Groups 2 and 3 completed Task 1. During week 3, groups 1 and 2 repeated Task 1 (using the same photo as they used for Task 1 the previous week). Then they completed a new task (Task 2 using a new photo). Group 1 then immediately transcribed their recorded Task 2 narrations to send to their teacher. Group 3 completed Task

2. During week 4, groups 1 and 2 (repetition groups) repeated Task 2, then completed a new task (Task 3 using a new photo). Group 3 completed Task 3. After each group completed Task 3, all groups immediately repeated Task 3.

Results revealed some positive effects that feedback and task repetition had on oral language performances in terms of CAF. In terms of repeated tasks, regarding complexity, group 1 (task repetition and post-task-teacher-corrected transcription) showed significant differences in amount of subordination and level of lexical variety over group 2 (task repetition group). There were no significant differences in mean length of AS units between the repetition groups. Regarding accuracy, there were no significant differences between the repetition groups in terms of increases in error-free clauses or improvements in verb form usage. Regarding fluency, group 1 showed a significant difference in speech rate over group 2.

In terms of CAF in the new task, regarding complexity, the three groups' performances in level of subordination and in mean length of AS units were significantly different, i.e., group 1 (task repetition and post-task-teacher-corrected transcription) over group 2 (task repetition group) over group 3 (control group). There were, however, no significant differences in lexical variety among the three groups. In terms of accuracy, there were no significant differences in increases in error-free AS units or error-free clauses. Regarding verb forms, each feedback group showed significant differences over the control group. In terms of fluency, there were no significant differences across the three groups.

In sum, results from many of the oral task repetition studies described above reveal that improvements of various CAF criteria improve at the expense of the others, and it is not consistent for each study's findings in terms of the same specific criteria in direct competition with the others. Another observation on the current body of empirical research on oral task repetition is that, so far, studies have looked mainly into university English language learners' narrative performances.

It is important to note that there are some differences in findings between studies, but that apart from participant, task and setting characteristics differences between studies, there were also differences in methodological designs and focus between studies, which might also explain differences in the findings. For example, the studies reviewed above varied in terms of exact versus procedural repetitions, task-type repetition, number of repetitions, etc. In fact, quite a few studies represent an initial task performance and one task repetition (so a total of two performances), thus, there is scope for more research that looks into more than 1 repetition (so a total of three or more performances).

It is also the case that the majority of task repetition research has looked at effects on oral performance, not written. Khezrlou (2020) and Manchon (2014) state that in consideration of the predominance of "oral modality in TBLT studies on the whole and [task repetition] studies in particular, it seems essential to expand the existent accounts of [task repetition] with due attention to writing tasks" (Khezrlou, 2020, p. 32). This helped inspire me to examine effects on students written performances after repeating tasks. The next section delves into studies on written task repetition and CAF.

2.3.5 Written task repetition and CAF

While many of the studies on task repetition conducted so far focused on oral L2 production, some research focused on the effects of task repetition on written language. The rationale is based on the idea that written task repetition provides the ability to free up students' limited attentional resources that help them devote more of their cognitive resources to properly structure aspects of language (Ahmadian & Tavakoli, 2011; Ellis, 2005). Compared to the higher number of studies on oral task repetition and CAF, there are some studies on written task repetition and CAF (e.g. Jung, 2013; Khezrlou, 2021; Indrarathne, 2013). Below, I review several studies on the effects of written task repetition on CAF performances.

Larsen-Freeman (2006), an early study, conducted a task repetition study on five EFL students in China, aged 27-37, without a control group or feedback conditions. She asked the students to write a narrative about a past episode they wanted to share. They carried out this task, no time limits, four times over a six-month period, with six-week intervals between performances. Then for each performance, three days after writing, they told the story orally. Across the six-month period of this study, written complexity, accuracy and fluency increased, though it can be argued that it was a combination of both written and oral repetition effects. One limitation of this study is it did not control for the possibility that CAF increased partly because of general language learning across the six months.

Azizzadeh and Dobakhti (2015) conducted a written task repetition study with 40 high-intermediate EFL learners in Iran, aged 18-25. Over 14 weeks, two groups (repetition versus no repetition) completed narrative writing tasks based on wordless picture stories. To start, the Nelson 300D test of English homogeneity was used to verify similarity in language proficiency scores. The repetition group repeated the same tasks one time, two to three weeks after initial performances. Pre- and post-tests (narrative writing task based on the same two sets of wordless picture stories) were used to compare writing levels among the groups. The findings

indicated that the experimental group significantly outperformed the control group in improved complexity. However, there was no significant effect in accuracy.

Next are two studies that looked at the effect of repetition on written performances for same-task versus procedural task repetition groups.

Jung (2013) conducted a study on written task repetition and the effect of feedback with eight Korean ESL students. They were placed in four groups: same task repetition with feedback; same-task repetition without feedback; feedback without same-task repetition; no repetition, no feedback, so a different task for the two "no-repetition" groups in which case the no-repetition groups would be procedural repetition groups. All participants completed essays from TOEFL prompt lists within 30 minutes. Two groups (same-task repetition) wrote an essay about living in big cities versus small towns, then repeated the same task, same topic, again one week later. The no-repetition groups wrote an essay on Day 1 about whether they preferred to wake up early, then on Day 3, they wrote about a different topic, their preference about living in a large city versus a small town. The feedback groups received corrective feedback two days after the first performance, and then were asked immediately after to spend 15 minutes to revise their writing without access to the feedback on first performances. The no-feedback groups were asked to spend 15 minutes to review their essays on their own before revising them. Seven days after the first performance, the repetition groups repeated the same task as they completed Day 1, while the no-repetition groups wrote a different task. Then two days later, the feedback groups received corrective feedback, and then were asked immediately after to spend 15 minutes to revise their writing without access to the feedback as they worked on their second performances. The no-feedback groups were asked to spend 15 minutes to review their essays on their own before revising them.

The findings indicated that all groups showed improvement in accuracy, but feedback conditions had limited effect. The repetition groups showed improvements in fluency and complexity, but the no-repetition group, i.e., different topics, slightly decreased in complexity. The groups that wrote on different topics improved in fluency to a lesser extent than the repetition group.

Next is Amiryousefi's (2016) study that, like Jung (2013), looked at same-task versus procedural repetitions. The main purpose of Amiryousefi's (2016) study was to examine the effects of task repetition versus procedural repetition on CAF for 70 low-intermediate Persian EFL learners' computer-mediated L2 written production. A secondary purpose was to investigate the relationship between computer anxiety and EFL learners' development of CAF in L2 writing. The participants were randomly assigned to one of two groups: (1) task

repetition and (2) procedural repetition groups. All participants completed a pre- and post-test. The task repetition group repeated the same task procedure with the exact same content five times; the procedural repetition group repeated the same task procedure with different content. The participants were also asked to complete the Computer Anxiety Rating Scale to measure their computer anxiety.

The results revealed that there was no statistically significant difference between the task repetition group and the procedural repetition group in terms of the pretest (Task 1) scores obtained on all CAF subdimensions. Thus, this result indicates that the groups were comparable in terms of complexity, accuracy, and fluency in L2 written production prior to receiving the treatments. Together, the results of statistical analysis in Amiryousefi's study revealed no significant relationship between computer anxiety and the participants' scores on all CAF subdimensions on both pretest and posttest. The task repetition group performed significantly better in terms of all fluency measures and in terms of percentage of error-free clauses (an accuracy measure), and the procedural repetition group performed significantly better in terms of the number of words and clauses (subdimensions of fluency). Findings from Amiryousefi's study help validate Kim and Tracy-Ventura's (2013) idea that both task repetition and procedural repetition can have beneficial effects on EFL learners' task performance.

In sum, results from most of the above studies show improvements in variations of CAF components through written task repetition. In the above findings, there was some variation of the specific CAF criteria that improved as a result of task repetition as well as the type of repetition. Together, repeating tasks generally results in improvements in L2 production. Like the results from oral task repetition studies introduced in section 2.3.4, the specific improvements in CAF criteria in terms of the ones in competition with one another varied based on each of the written task repetition studies. In terms of generalizations, there so far have been fewer studies on written task repetition than on oral task repetition, which leaves room for designing a study that investigates the effects of written task repetition.

2.4 Feedback in task repetition

The existing body of research in L2 learning emphasizes the importance of feedback as part of the task repetition process (Ellis, 2009; Manchon, 2014). Manchon (2014), for example, argues that "the availability of feedback and the role of feedback in bringing about potential benefits should be made central in future [task repetition] preoccupations… these gains are purported to

be crucially dependent on the learners' own engagement with and processing of the feedback received" (p. 31).

Feedback is hereby understood as the teacher's response to a student's performance. Such a response could align with Kulhavey (1977) who described feedback as "any of the numerous procedures that are used to tell a learner if an instructional response is right or wrong" (p. 211). However, the necessity for feedback for effective learning goes beyond identifying errors. It helps students understand new information and serves as guidance on how to improve (Bellon et al., 1991). Also, additional/alternative forms of feedback include teacher repetition and rephrasing of student L2 production, recasting, explicit correction, elicitation, clarification requests and metalinguistic feedback (Ferreira et al., 2007).

When it comes to providing feedback on writing, a major goal is to engage the students to revise their work through repeating the tasks. Ferris and Hedgcock (2014), for example, state that "[b]oth teachers and students feel that teacher feedback on student writing is a critical, non-negotiable aspect of writing instruction" (pp. 237-238). Though tasks are primarily meaning based, they offer opportunities for some language focus as well as feedback on meaning. Further, feedback on students' performances allows students to attend to and develop their L2 skills, thus making feedback important to incorporate into written task repetition (Ellis, 2009; Manchon, 2014).

Feedback on language focus or meaning can be very helpful for students even if they do not repeat the same task. In fact, many researchers have also argued that if learners can apply the knowledge from the feedback from earlier writing tasks to new writing tasks, then the results would likely suggest the effectiveness of "feedback for acquisition" (Manchon, 2011; Polio, 2012; Truscott, 2007). Potentially, that acquisition would also improve regardless if the next task is the same as the previous task, i.e., learners can still glean much of the feedback commentary by applying the more general feedback, i.e., general/overall comments, to new tasks. Several empirical studies on task repetition included feedback conditions that were described in their research, for example, Mennim (2003) and Jung (2013), sections 2.3.4 and 2.3.5, respectively. There are some studies on feedback and written task repetition where there were significant differences in student performances in terms of specific CAF components as a result of feedback conditions (e.g. Jung (2013); Roothooft, 2022; Kim et al., 2022; Kim & Li, 2024).

The following is a series of studies on feedback in task repetition research. Some of the task repetitions were based on the same task while other task repetitions were based on new tasks. Also, there are differences in the type of feedback given in different studies. The number

of participants furthermore widely varied between studies, yet what the following studies have in common is they were based in university English class settings, incorporating academic assignments that required summarizing and critical thinking. As is the case for the task repetition literature more generally, most studies have focused on feedback in the context of oral task repetition, with limited work available on feedback on written task repetition. As is the case for the task repetition more generally, most studies have focused on feedback in the context of oral task repetition, and to my knowledge less work is available to date on feedback and written task repetition (e.g. Khezrlou, 2019; Roothooft, 2022; Kim et al., 2022; Kim & Li, 2024).

In Mennim's (2003) oral task repetition study, the university students received feedback, and then oral performances improved at the next performance. However, without a control group, i.e., a no-feedback group, in Mennim's study, it is not certain the extent to which it was the feedback that helped the students improve their oral performances as opposed to the fact that they were in a university course, or simply the repetition. The likely reason for no control group was there were not enough participants., i.e. one group of three students. In Jung's (2013) study, a control group for no-feedback was included, thus enabling the comparisons to be made for feedback conditions.

As described in section 2.3.5, the findings in Jung's study indicated that all groups showed improvements in accuracy, although feedback had only a limited effect on accuracy. The repetition groups showed improvements in fluency and complexity, but the no-repetition group, i.e., different topics, slightly decreased in complexity. To be more specific, with fluency, the feedback group showed statistically significant improvements in fluency by performing more continuous writing whereas the no-feedback group showed only slight fluency gains. With complexity, the feedback group showed significant improvements in syntactic complexity, for example, more C/T whereas the no-feedback group showed very little improvement as they repeated the tasks. The groups that wrote on different topics improved in fluency to a lesser extent than the same-task repetition group.

A rare study conducted in the context of integrated writing tasks is Kim and Kim's (2017) investigation of the impact that feedback had on the written performances of repeated integrated reading-write tasks that ten Korean MA TESOL graduate students at a university in Korea (aged 20-30 years old) completed during a seven-week period in a research methodology course. Kim and Kim's study looked at the way students used teacher feedback on the content component in their revisions, i.e., to gauge the effectiveness of feedback on

writing performances as well as to gather student perceptions in terms of the type of feedback that they felt helped them the most.

Each week, the students each independently wrote a one-page review based on an academic assignment, scholarly journal article (a different topic for students to choose from each week), where they were asked to summarize the article and provide a critique (without a time limit). Then, within the same week, all participants received teacher feedback on the content of their writing, i.e., the article summary and critique, not feedback. The participants received mainly indirect feedback, for example, the teacher wrote "Provide more details, What do you mean by this?, This is a great summary, Can you make a better transition here?" (p. 62), and a minimal amount of direct feedback, for example, where the teacher made a correction or stated that an error was made and recommended a revision. The teacher provided only comments and not numerical scores out of concern that "some students may focus only on the numeric scores rather than the written feedback if provided with both" (p. 61).

One week later, the students submitted their revisions as an assignment, and because they wrote their revisions outside of class, they were allowed access to their first drafts with the feedback when doing the next drafts. During each of the following weeks, they completed their next task which entailed summaries and critiques of journal articles. However, each student's initial writing was always based on a new article. Since there was a new task each week, this study examined whether students applied the feedback not only on the revisions but also when they completed subsequent summaries using the same format as the previous tasks but based on new articles.

This study includes an interesting feature that shows how the teacher gauged the students' uses of the feedback. In terms of the revised tasks (second written performances of the previous week's tasks), Kim and Kim used a 0-5 point scale (0 = no evidence; 5 = excellent) to compare the first and second performances to the extent to which the students addressed the teacher's feedback: addressed, partly addressed, and not addressed. The results showed that overall, there were improvements from Task 1 to the next four subsequent tasks (Tasks 2, 3, and 4), but dropped at Task 6 to a somewhat similar score as Task 1. Not all student scores showed that they used the teacher's feedback; some student scores fluctuated in the sense that some decreased or remained the same while several students' scores increased over time.

Kim and Kim (2017) then interviewed two of the students (Lee and Kim) in a semi-structured format to gather their perceptions about the effectiveness of teacher feedback, with a particular focus on the type of feedback they found most beneficial for improving their writing

as well as the type(s) of feedback that they would like to receive after completing future assignments. Lee, whose writing scores improved over time, indicated that the feedback helped her improve in selecting the most essential information from an article for her to use when summarizing the reading. Also, she stated that the feedback helped her make her article critique less ambiguous so that the reader knows her specific viewpoints. Kim, whose writing scores remained the same over time, indicated that the writing activities helped her become more confident in writing article reviews. She stated that when the feedback indicated "provide more details," it helped her identify with the reader's perspective in mind when she completed subsequent reading-write tasks. Kim also suggested that peer assessment be incorporated so that she receives even more feedback, thus providing more opportunities to improve before submitting the next task completion. Both students reported that they prefer indirect feedback over direct feedback because the indirect method, in turn, requires them to apply their critical thinking skills more than direct feedback would. Although this interview on feedback is based on a very small sample, it is worth noting that even when a student's writing score does not improve that it is still possible that the student might still state that they found the feedback helpful. It would be informative in future studies to gather student perceptions on feedback including from students whose writing scores decreased over time, an opportunity to make comparisons between students' writing when comparing the feedback and no-feedback groups. Additionally, a larger sample could help confirm the positive impact that task repetition and feedback have on student writing as well as the positive perception that students have of teacher feedback.

In sum, the findings from the empirical studies discussed in this section demonstrate that feedback is helpful in various aspects of students' language performances. It is hoped that results from the present study on task repetition that also investigates the impact that feedback has on language performances will show similar positive findings, thus an encouragement for teachers to continue to provide feedback. There continues to be room for research on ways that feedback helps learners improve their language production, which is what has inspired me to incorporate feedback condition as the treatment intervention group in my study.

2.5 Perceptions of task repetition

As shown from the results from a number of the empirical task repetition studies discussed in earlier sections of this chapter, task repetition has played a role in helping learners improve in their language performances. However, some researchers have argued that repeating the same task is susceptible to monotony and boredom (Ellis et al., 2019; Ahmadian et al., 2017).

Although there is much evidence confirming the benefits of task repetition for speaking, Larsen–Freeman (2018) and Ahmadian et al. (2017) have argued that some educators are reluctant to incorporate it in class out of concern that learners would find task repetition boring. Because classroom studies in general suggest that repetition tends to induce boredom, several researchers (e.g. Geiwitz, 1966; Tavakoli & Hunter, 2017; Kruk & Zawodniak, 2020; Hanzawa & Suzuki, 2023) have called for further research to investigate how this practice is perceived by second language (L2) learners.

Ellis et al. (2019) stated that "exact repetition, which requires learners to repeat the same task multiple times, is not the ideal option" (p. 231) because some learners would be discouraged by the monotony. In fact, as far back as 1993 Plough and Gass had already expressed the concern that when a student becomes familiar with a task, then the originality of that task fades, and thus becomes boring.

To explore the relationship between task repetition, language performances, and students' perceptions of repeating a task, researchers and educators must ideally capture the students' perceptions of task repetition. Next are empirical studies that examined students' perceptions of task repetition.

Aiming to shed more light on the perception of boredom resulting from task repetition, Ahmadian et al. (2017) interviewed eight L2 teachers regarding their views on incorporating task repetition in classes. Although most of these teachers stated that they were aware of the language performance enhancement resulting from task repetition, they also indicated that learners would likely become bored with the task and lose interest in engaging with the work. Ahmadian et al.'s (2017) study investigated ways that language teachers and students perceive oral task repetition. This study consisted of eight language teachers and 21 intermediate to upper-intermediate level language learners (20-24 years of age), in a language school in Iran. The learners had already completed 10 months of English study. These learners were university students who would eventually take the TOEFL or IELTS examinations to then apply to universities abroad. The task used in this study was an oral narration of a picture description. The students were assigned to pairs to discuss how they plan to tell a story about the picture to their classmates. Each pair was given two minutes to discuss their storytelling plan, but they were not allowed to take notes. Then, immediately after the two minutes, they narrated their stories to their classmates (no time limit on narrations). The students then repeated the same task one week later in different pairs.

After the students completed the task repetition, the students and teachers immediately participated in one-to-one semi-structured interviews, seven to nine minutes long, that the

researchers recorded. In terms of student interviews, the researchers inquired about ways students perceived task repetition, its purpose, and the way(s) they felt that task repetition impacted their language performances. In terms of teacher interviews, the researchers inquired about the ways they anticipated students would perceive task repetition as well as the facets of language production they expected task repetition would most likely impact.

Examples of questions that students were asked included "How do you feel about repeating a task?", "How do you think repeating the same task affected your performance?", and "What strengths and weaknesses do you see in task repetition" (Ahmadian et al., 2017, p. 6). Regarding task repetition, results showed that 18 out of 21 students responded with statements similar to "Repetition shows consolidation of what we already know", "…although initially I did not know what the purpose of task repetition was, it helped me a lot in that I could state the [same] sentences…in a more fluent, organized and accurate fashion", and "in the first performance, I struggled to figure out what [to say]…but this time I…could repeat the same content with a better sentence" (Ahmadian et al., 2017, p. 6). Regarding the aspects of language production that task repetition impacts, results showed that the majority of students (16 out of 21) perceived task repetition to improve fluency and accuracy. Some examples of student responses included "I was quicker on the second occasion", "I could speak faster", "I did not have too many pauses", and "I am pretty confident that I used [the word assignment] in the right sentence…could use more accurate sentences" (Ahmadian et al., 2017, p. 6). Regarding perceptions, the majority of students (18 out of 21) indicated that they did not find task repetition boring while the other three students suggested that the subsequent tasks be slightly different than the previous tasks. This positive indication that students do not view task repetition as boring aligns with a similar finding from Lambert, Kormos, and Minn's (2016) study that looked at the relationship between oral monologue task repetition and immediate L2 fluency gains.

Examples of questions that teachers were asked in Ahmadian et al.'s research included "What do you think your students think and feel about task repetition?", "What aspects of language do you think task repetition is more likely to affect", and "What strengths and weaknesses do you see in task repetition?" (Ahmadian et al., 2017, p. 9). Results showed that the majority of teachers (6 out of 8) believed that task repetition impacts only fluency, unlike the majority of students who believed that task repetition impacts fluency and accuracy. Similar to students' predictions, all teachers felt that task repetition is an effective teaching practice that fosters enhanced L2 proficiency and use. For example, most teachers found that task repetition improves student confidence in completing subsequent tasks. Some examples of

statements that teachers made to support these results include “if learners do a task once they will have more self-confidence to do it a week after” and “…the first performance leaves some traces in learners’ memory and therefore they may be able to do it more effectively [next time] because they know more about the content” (Ahmadian et al., 2017, p. 6). The similarities about teacher and student perception of the effects of task repetition, however, stop here. While the majority of students stated that task repetition is not boring, the majority of teachers (7 out of 8), as introduced earlier in this section, believed that students would find it boring and would lose interest in engaging with the work. Potentially, because the teachers’ anticipated views that students have about task repetition did not always align with student responses about task repetition, teachers and researchers need to investigate more closely about ways that various student populations perceive repeating tasks. It is equally possible that teachers do not necessarily hold accurate views on their students’ perceptions.

Another study that examined student perceptions of oral task repetition is Hanzawa and Suzuki’s (2023). Some findings from earlier studies, such as Ahmadian et al.’s, suggested that teachers are reluctant to incorporate task repetition because of their concern that students would have negative perceptions about it. Hanzawa and Suzuki’s study looked at students’ perceptions of oral task repetition in relation to emotional engagement and metacognitive judgement. More specifically, their study investigated the number of repetitions of the same task that students find most helpful (metacognitive), the extent to which three different time intervals between performances impact metacognitive judgement and emotional engagement, and the degree to which metacognitive judgement and emotional engagement relate to fluency improvements across three groups of students that each had different time intervals.

Hanzawa and Suzuki’s study comprised 64 university EFL students at a university in Japan. The students had been studying EFL about six years before starting university studies. TOEIC scores were collected to determine whether there were significant differences in scores among the 64 participants within this pool. After it was determined that they had similar proficiency based on these scores, they were split somewhat evenly among the following three groups (conditions): group 1 (massed-spaced, n=20); group 2 (short-spaced, n=23); and group 3 (long-spaced, n=21). All students completed the same picture description narrative task six times under their groups’ assigned time interval conditions. Group 1 (massed) had no time intervals between performances; group 2 (short) performed the narration three times in a row, then were given a forty-five-minute time interval before completing the next three performances one after the other. Group 3 (long) completed three narrations in a row, then one

week later, they completed the other three narrations one after the other. Groups 2 and 3 participated in regular class activities after the first three performances.

This study took place in a computer lab where the students worked independently where they first viewed their assigned six-frame cartoon pictures as they listened to the researcher's pre-recorded story narration two times. Then the students were provided with 12 useful vocabulary words. Next, they were given 90 seconds to plan ahead on how each would subsequently perform a two-minute narration that was recorded into the computer software as they spoke. During the narrations, they had access to only the pictures. They were instructed to begin their narrations with "Yesterday I saw an unusual event."

After each group completed their sixth narration, they were provided with an online questionnaire that probed into their perceptions about task repetition. Some examples of statements that probed directly into student perceptions included "I would do this task again", "I was bored doing this task", and "This practice excited my curiosity" (Hanzawa & Suzuki, 2023, p. 23). Other questions in the questionnaire asked students the maximum number of performances they felt were needed for their L2 performances for improvement. They were also asked what aspect(s) of their performances they felt had improved.

In terms of the results, regarding fluency, the results showed that for the first three task performances, for all groups taken together, the mean length of narration was similar at each performance.. At times 4, 5, and 6, the long-spaced group, in terms of fluency, showed significantly lower performance than the other groups. In terms of students' perceptions of task repetition, all participants in the study indicated on the 7-point Likert scale that task repetition was beneficial to improving L2 proficiency (mean score 4.66 out of 7), and survey results showed that producing four to five performances was perceived by the students as most effective.

Like Ahmadian et al.'s (2017) study, there were very positive student perceptions of task repetition. There were no significant differences between the groups in terms of their positive views on the effectiveness of or desire for repeating tasks. This suggests that students, overall, found task repetition helpful.

The findings from the above studies exemplify that task repetition can be an effective pedagogic tool, at least with reference to oral task repetition. To my knowledge, no studies are available that have looked at student perceptions of listening-to-write task repetition. Thus, research on this latter topic seems substantiated.

2.6 Integrated tasks

Writing has often been the sole construct in writing courses, and such independent assessment tasks have been a key part of academic success (Kellogg & Raulerson, 2007), and has often been used as a component of any second language (L2) proficiency assessment tool used for academic admissions decisions (Chapelle et al., 2008; Taylor & Angelis, 2008). Even though many of those assessments are independent tasks based on personal experience in the 'real world', independent writing tasks do not represent the majority of writing tasks that students will encounter in academic settings, thus lacking validity (Cumming et al., 2005; Read, 1990; Weigle, 2004). Writing is often not an autonomous language skill but rather combined with reading, listening, and/or speaking, and various tests have incorporated integrated writing tasks in addition to independent tasks (Biber et al., 2017; Deluca et al., 2013).

To that end, an increasing number of language proficiency assessment tools include integrated writing tasks, for example, an essay response based on information provided in a listening and/or reading passage (Kyle, 2020), which relate more closely to authentic academic contexts (Chapelle et al., 2008; Cumming et al., 2005). For academic and professional success, this requires the ability for learners to glean ideas from auditory, visual, or mixed-media sources and then incorporate them into complex original writing (Abrams, 2019; Cooney et al., 2018; Plakans, 2010, Plakans & Gebril, 2013). Integrated tasks such as reading-to-write or listening-to-speak tasks are therefore increasingly used in EAP instruction and L2 assessment (Rukthong, 2016; Rukthong & Brunfaut, 2020). An example of a test which contains a task that requires test takers to listen to and read a dialogue and then summarize the content in writing is the Test of English as a Foreign Language (TOEFL). Other high-stakes tests that use integrated tasks are, for example, Pearson Test of English (PTE) Academic, Canadian Academic English Assessment (CAEL), Ontario Test of English as a Second Language (OTESL), and Certificate of Proficiency in English (COPE) (Yang & Plakans, 2012).

Authentic materials that integrate multiple language skills are especially important for EAP classes because they address communicative purposes that are needed in an immersion environment that provides a realistic context for tasks that relate to learner's needs (Benavent et al., 2011). Because such tasks should align with the real world, they should not be created primarily for pedagogical reasons (Benavent et al., 2011). Integrated tasks can help prepare learners to address skills in problem-solving, project-based learning, case-based learning, role-play, and simulation. Students, teachers, and test developers can use authentic materials as a means to "link the formal, and to some extent artificial, environment of the classroom with the real world in which we hope our students will eventually be using the language they are learning" (House, 2008, p. 53). Examples of EAP tasks with authentic academic activities

include taking notes in a meeting or lecture, doing a presentation with a Q&A session, and reviewing a book or film.

Rukthong and Brunfaut (2020) explain the advantages of integrated task usage in assessment contexts: (1) Better authenticity than short-answer tests; (2) real-life language use outside testing situations; (3) allows test-takers to generate content of responses from the input; and (4) positive washback in the classroom. According to Plakans et al. (2018), connecting writing with another language skill improves not only each language skill but also academic L2 performance more generally.

My study takes place in an academic context; therefore, it is relevant to consider the nature of language tasks in such contexts. Academic ESL writing tasks are usually integrated with at least one other skill to elicit more authentic integrative language use (Plakans, 2009; Hinkel, 2006). According to Plakans (2010), language test developers and educators need a greater understanding of how writers respond to integrated tasks as well as compose meaning.

In spite of the prevalence of integrated writing tasks on high-stakes academic tests, there are concerns raised about the extent to which performance is impacted by learners' listening and reading comprehension abilities (Payant et al, 2019). Also, if learners do not understand the aural and written text input or have difficulty identifying important ideas and concepts, they may not perform well on integrated tasks. Payant et al. (2019) stated: "Integrated writing tasks require more than basic comprehension because the source information must be interpreted and reinvested into students' own texts" (p. 88).

Research on integrated tasks is on the increase, and in particular, several studies have focused on reading-to-write tasks (Shin et al., 2015; McCulloch, 2013; Ohta et al., 2018). Interest in reading-to-writing integration began in the early 1980s in which researchers positioned that reading and writing "share similar composing practices and thus should be treated jointly, not separately, a view that was the driving force behind the reading-writing connections work that appeared in the 1980s and remains a useful framework today" (Manchon & Matsuda, 2018, p. 573). As other reading-writing connections were researched, results from studies suggested that this became a critical link within the field of EAP where the focus is on preparing students for the literacy demands within the context of the courses that they would take.

There are many studies on integrated tasks (e.g. Zhao et al., 2024; Yang & Plakans, 2012; Plakans et al., 2019; Payant et al., 2019; Plakans & Gebril, 2012). Next, I discuss a series of empirical studies that looked at various dimensions of integrated writing tasks where students need to synthesize information that they gleaned from language input to produce their

written performances on the integrated tasks. The selection of studies below provides various examples of integrated writing task research. Many of the integrated tasks are derived from high-stakes English language examinations, for example, TOEFL, Pearson Test of English Academic (PTE), Canadian Academic English Language (CAEL) Test, and other high-stakes examinations that are used in EAP settings.

The following are four empirical language studies that used integrated writing tasks, of which the first three studies described are integrated reading-listening-writing tasks.

Yang and Plakans (2012) used an integrated reading-listening-writing task to explore L2 learners' writing strategy use in terms of the mental and behavioral activities relevant to the before, during, and after writing stages of the written test. This study consisted of 161 non-native English-speaking university students in the United States (47 undergraduate, 88 graduate, 26 non-matriculated). Their English proficiency varied based on TOEFL scores.

The reading-listening-writing task used in this study came from the TOEFL iBT Data Set 3: Writing test. The students had two minutes to read a 255-word passage, then they listened to a two-minute pre-recorded lecture that pertained to the content from the reading passage. Students were allowed to take notes during the listening input. They were then given a twenty-minute time limit to plan and write a 150-225-word response to the prompt. The written part of the task consisted of a question that asked students to summarize ideas from the lecture (listening input) and to discuss the relationship between those ideas and the content from the reading. The writing task was scored based on content, organization, language use, and degree of verbatim source use. Before the writing test began, the students were informed of the expected word length and time limit.

After the students completed the test, they completed an online Standard Strategy Inventory for Integrated Writing (SSIIW) questionnaire. This questionnaire consisted of thirty-four statements on a 1-5 Likert scale (very rarely to very often). These statements probed into "mental and behavioral activities related to specific stages (before, during, and after writing) in the process of completing an integrated reading-listening-writing task" (p. 85).

Findings from Yang and Plakans' (2012) study suggest that the strategy that positively correlates with written performance is discourse synthesis strategy, where students select, organize, and connection information that they learned from the language input when producing their language performance on the integrated task. This correlation connects with quality writing in terms of language use, content, and organization. Conversely, when students use shortcuts such as copying information based on verbatim memory from the input, this strategy, test-wiseness strategies, negatively correlates with the writing quality. This suggests

the need for students to practice strategies, most notably discourse synthesis strategies, as they complete an integrated writing task which would require them to choose ideas from the input, organize them logically, and connect the ideas.

Zhu et al. (2016) used an integrated listening-reading-writing task to examine the relationships in students' performances between an integrated listening task versus an independent listening task and to investigate language competencies in the integrated writing test. This study consisted of 226 native Chinese Secondary Five students, an average age of 17 years, from six Hong Kong secondary schools. The design of this task is similar to an integrated writing assignment of Chinese Language in the Hong Kong Diploma of Secondary Education (HKDSA) to assess student writing based on their comprehension and use of listening and reading materials (input).

The integrated listening-reading-writing task in this study lasted for 60 minutes. First, students listened to a pre-recorded audio conversation between two students who disagreed on the specific landscape that they would like to preserve at their school. After they listened to the recording, they then read five texts (about 2400 Chinese characters total) that pertained to various aspects about the refurbishment plans and the school's history. After that, the students were asked to write an article (400 words minimum) that would show their views through the lens of one of the two students from the recording regarding their preferred type of school landscape that should be preserved.

The independent listening task in this study was a 30-minute task that consisted of two recordings, both based on the same topic as each other. After the students listened to the recordings, they were asked to complete a 14-item test (nine multiple choice and five short-answer questions). This independent task assessed the following six types of listening comprehension processes: (1) memorization (retelling information); (2) explanation (paraphrasing); (3) summarization (sorting ideas and relating them to the context); (4) elaboration (making inferences); (5) evaluation (critical thinking); and (6) creation (problem solving and offering own viewpoints). The first three processes require a basic understanding from the input; the last three processes require students to go beyond merely demonstrating a basic understanding of the input, i.e., advanced critical thinking skills.

Results from Zhu et al.'s (2016) study showed statistically significant correlations between the students' performance on the independent and integrated tasks. The basic comprehension level indicators, i.e., memorization, explanation, and summarization, of the independent task did not show significant correlations with the indicators of the integrated task. However, the elaboration, evaluation, and creation processes (beyond the comprehension

processes) of the independent task were significantly correlated with many indicators of the integrated task, for example, tone, interaction, synthesis, language use, organization, etc. Because of this significant predictability of scores between the independent and integrated tasks, these findings suggest that higher-order skills from Bloom's Taxonomy are instrumental (as opposed to lower skills such as memorization, explanation and, summarization) for success in integrated task performances.

Plakans et al. (2019) conducted a study to investigate integrated writing assessment performances with respect to the linguistic features of CAF. Because integrated tasks in large-scale classrooms are used, validity evidence was needed to support the claim that the scores reflected specific targeted language abilities. Four hundred eighty integrated TOEFL test performances were analysed using CAF measures to determine the extent to which these linguistic features could predict scores on reading-listening-writing tasks. Plakans et al. (2019) used TOEFL rubrics as part of their analytic tools. In a review of the scoring rubric for TOEFL integrated writing for CAF, Plakans et al. reported that accuracy is found to be the only one of the three features that is overtly addressed in the rubric through phrases such as "occasional language errors" and "errors of usage and/or grammar". Complexity and fluency were connected to criteria in the scale related to "imprecise presentation" and the inclusion of main ideas from the source texts. However, the rubric does not directly mention either fluency/development or complexity/sophistication.

Results indicated a cumulative impact on scores from the CAF measures: Fluency was found as the strongest predictor of integrated writing scores; for accuracy, analysis of linguistic errors revealed that morphological errors contributed more to the regression statistic than did syntactic or lexical errors; complexity was significant but represented the lowest correlation to score across the variables. It is worth noting that studies by Cumming et al. (2006) and Plakans and Gebril (2013) provide foundation for Plakans et al.'s (2019) study except that only two rather than three skills were required to complete this integrated reading-listening-writing assessment.

So far, many reading-writing and listening-speak integrations have been investigated, but fewer listening-writing integrations have been researched as instructional or assessment tools for language learners (Manchon & Matsuda, 2018). Where listening-writing tasks were investigated, often these incorporated an additional input skill, for example, reading-listening-writing. The present study, however, aims to gather insights on a task that integrates only listening and writing. I next describe examples of empirical studies on integrated reading-listening-writing tasks, then a listening-writing task that were used.

Zhao et al. (2024) conducted a study using an independent writing task and an integrated reading-to-write task to investigate the effects that independent versus integrated writing tasks have on EFL learners' cognitive demands in terms of linguistic complexity, i.e., lexical and syntactic. This study comprised of 35 undergraduate Chinese EFL learners who were English majors in their second year of university study. The independent writing task was for students to write commentary about their views regarding whether there were positive or negative impacts on people's lives in terms of social networking. The integrated writing task was for students to begin by summarizing a 270-world news article about environmental effects caused by plastic. Then, the students were asked to write their suggestions about ways to increase public awareness about saving the environment through a plastic bag ban. This two-part computer-based integrated writing task was set with a 200-word minimum for each part.

After each of the two writing tests, the students completed self-perception questionnaires where they used a 100-point scale to rate their perceived level of difficulty of the task as well as the mental efforts they feel they exerted for each task completion. To measure linguistic complexity, computerized complexity analysis software was used. For syntactic complexity, measures included T-unit, length of text, coordinate phrases, and phrasal sophistication. For lexical complexity, measures included lexical diversity and lexical sophistication.

Results from Zhao et al.'s (2024) study revealed that cognitive demands were placed significantly higher from integrated writing tasks than from independent writing tasks. Also, the integrated task performances had a higher degree of lexical sophistication and syntactic complexity, i.e., more advanced sentence structures and longer sentences. Findings from this study suggest that when tasks are more demanding, i.e., requiring more cognitive skill practice, students are able to write more complex language output.

Having reviewed studies with reading-listening-writing integration, as well as a reading-to-write integration, I now give an example of a study on integrated listening-to-write tasks (in fact, also including listening-to-speak tasks) that did not integrate reading input in the following empirical study.

Rukthong (2016) used integrated listening-to-summarize tasks (listening-to-speaking and listening-to-write), adapted from the Pearson Test of English (PTE) Academic, to investigate how students process the listening input. Seventy-two EAP students (Thai-L1, 20-40 years of age) who were pursuing various majors (mostly postgraduate) at different universities in the United Kingdom participated. A total of eight listening-to-summarize tasks were investigated in the study – four listening-to-speak and four listening-to-write. In all tasks,

students first listened to a 60-90-second prerecorded lecture. At the same time, a picture that pertained to the listening input was provided and the students were allowed to jot down notes during the lecture to then use for the rest of the test. For the listening-to-speak task, they were given 10 seconds to prepare a speech, then they were given 40 seconds to retell in 50-70 words what they heard in the lecture. For the listening-to-write task, before the students heard the listening input, they were informed that they would first hear the lecture and then would be given 10 minutes to prepare and write a summary as if they were summarizing for a classmate who missed the lecture. Additionally, a questionnaire was administered, comprising of 23 statements on a 5-point Likert scale, that probed into students' perceptions about task authenticity, fairness and difficulty of each task.

To examine task difficulty, 60 students completed perception questionnaires. This questionnaire probed into the students' perceptions of the tasks and task difficulty. More specifically, the perception questionnaire probed into students' perceived task authenticity, fairness, and level of difficulty. The remaining 12 students conducted a stimulated recall to capture their real-time listening and summarization processes to investigate cognitive processing and strategy use during task completion. All students completed the questionnaire for each task to shed light on task perceptions.

Results from Rukthong's (2016) suggest that the majority of the students perceived the tasks as having been fair and authentic. In terms of the listening-to-summarize tasks, the students perceived them to represent real-world academic listening, thus this suggests student support for the use of these tasks for assessment. Both higher-level cognitive processes, e.g. drawing inferences, and lower-level cognitive processes, e.g. parsing, were used by the students. Results also suggest that a students' perceived difficulty is not always a predictor of student performance, i.e., a student's background knowledge and individual processing strategy can influence the degree of task difficulty.

In sum, these studies investigated various dimensions (e.g. cognitive demands, skill interaction, etc.) of integrated writing tasks which require students to synthesize information from listening and/or reading inputs to produce their writing performances. Because my study is positioned within a university, my writing task will be integrated (listening-to-write), not independent (listening comprehension only), to be in line with academic settings and assessments where tasks typically integrate more than one language skill. Given the scarcity of research on listening-to-write integration, as shown above, my focus will be on this task type.

2.7 Knowledge summary and transfer

Earlier in this chapter, I reviewed the literature on CAF with a focus on writing development. However, given the nature of my listening-to-write task, I will now introduce the concepts of knowledge summary and knowledge transfer as potential ways of operationalizing the integration of listening with writing. These are also relevant skills to the EAP context. Knowledge summary is about the extent to which someone understands the input; knowledge transfer is about the extent to which someone takes their understanding of the input further.

As for ***knowledge summary***, summarizing is a complex activity that is a highly essential and necessary skill, and summary writing is one of the most difficult for learners to master (Lin & Maarof, 2013). It is a major skill in academia because it is highly useful to digest large amounts of academic input and is associated with both reading and writing (and listening if the input is oral, e.g., summary of a lecture). It contributes to academic success, and it requires learners to relay ideas other than their own (Namibiar, 2007; Johns, 1985). If the input for the task is listening, then summarizing a lecture is an example. Summarizing also requires learners to differentiate between main ideas and supporting ideas from a passage, a skill on which many students fall short (Othman, 2009), thus a necessity to incorporate into L2 writing instruction.

Rinehart and Thomas (1993) state: "Writing an effective summary requires reflection and decision-making. They discuss how to relate text ideas, how to narrow important information to the level of organizational gist, and how to capture the gist in written form" (p. 24). Hidi and Anderson (1986) add that summarization requires learners to use existing text to apply their comprehension of the discourse already provided, thus a gauge of their skill to select ideas to compose a summary.

In order for learners to succeed in knowledge summary formation, excellence in listening and/or reading skills is critical. Learners must identify the most important ideas from a text or lecture, and this also requires their ability to centralize ideas while disregarding irrelevant details that the input might provide. Part of successful summarization, therefore, is providing condensed overviews of the input, thus providing a reader or listener in a short amount of time the gist of what was discussed/written. An advantage of reading over listening is reading provides visual information that learners can refer back to as they simultaneously absorb information and meaning. However, listening requires learners to process the input as they learn meaning in instant real-time as opposed to in reading which allows for more processing time, hence a strong need for listening to be strongly emphasized when incorporating knowledge summarization skills.

Some existing EAP assessments assess some knowledge summarization. For example, even though most TOEFL listening items are not on summarization, two of the constructs for the TOEFL ibT Listening Section are identifying main points and selecting details. Ability to identify main points is necessary to write a summary. While selecting details is not needed to include in a summary, the skill of differentiating main points from details is necessary for condensing the writing with salient points needed for summarization.

Next are some task repetition studies on listening comprehension **(knowledge summary)** in which materials from major assessments were used. In Sakai's (2009) task repetition study among 36 university EFL students in Japan who were mostly in their second or third year, the study looked at whether repetition affected students' ability to recall the content from listening input and whether the effect of repetition on listening comprehension was the same for students at different language proficiency levels. The students had six years of English language instruction before they started university study. The students were split into two listening proficiency groups based on the mean scores of the Michigan English Placement Test that they took the month before this empirical study, i.e., higher listening proficiency group and lower listening proficiency group. The listening portion of the Michigan English Placement Test was used for this study. The students listened to the first passage, about 60 words long, and were told not to take notes while the passage was being played but that immediately after, they would be given three minutes to jot down everything that they understood from the input. Then, immediately after, the second passage was played, the process was repeated. After that, the students were asked to put away their notes, and then the same process for the first set of repetitions was repeated. Subsequent to that, the students were asked to write what they retained from the input. Listening comprehension was assessed based on the students' free writing, i.e., Sakai counted the number of correctly recalled idea units from the original input. Results from Sakai's study revealed that both groups improved to a similar degree at Time2. There was a large main effect of repetition and listening comprehension for both groups. However, there was no interaction effect of time and proficiency level. These results suggest support for the use of repetition for developing learners' listening comprehension regardless of the students' levels.

While Sakai's (2009) study focused on a single activity for the listening test, i.e., jotting down notes, next is another university EFL task repetition study that used two types of listening comprehension tasks: multiple-choice questions and one open-ended question. In Iimura's (2006) task repetition study, 54 university EFL students in Japan were investigated on the effects of different task uses on listening comprehension and task repetition. This study

investigated the type of task on which listeners perform better, multiple-choice or open-ended; whether there were effects of repetition for either test; and whether there were differences in scores on the two tests in relation to the first, second, and third performance of each test. The students were randomly split into two similar groups. The multiple-choice test was extracted from the STEP Eiken Test in Practical English Proficiency, and the open-ended task used the same questions from the multiple-choice test with the choices removed. The passages, about 60 words in length, were selected from the listening part of the STEP test. Group A was given the test in multiple-choice format, Group B in open-ended question format.

A total of 10 passages were played three times with a 20-second time interval between each play. After each play, the students repeated the same items on which they wrote their answers. These steps were repeated for all 10 passages. Results showed that the mean scores for multiple-choice tests (Group A) were significantly higher than for open-ended tests. Results further showed that there were significant differences within both groups' scores depending on the repetition: for Group A, there were significantly different scores between the first and second performance, and between the first and third performance. For Group B, there was a significant difference between the first and third performance. These results suggest that for the open-ended test, unlike the multiple-choice test, there needed to be an extra repetition for there to be an effect of repetition.

Given the favorable results of repeating a listening task, a strong plea for listening in everyday instruction would be in order because listening is the most frequently used language skill, and it plays a significant role throughout the educational process and daily communication (Abbassian & Chenabi, 2016; Nunan, 1997; Brown, 2001). Listening is an important skill in daily communication, so it makes sense to teach it. When we teach it, we should use repetition as a pedagogic technique because it has been shown to be effective. Nunan (1997) states "listening is the basic skill in language learning. Learners will never learn to communicate effectively in the absence of an effective instruction, which assigns a pre-requisite role for listening" (p.47). Even earlier on, Lund (1990) supported that listening has become even more important for adult ESL learners because many curricula had become more extensively comprehension-based.

Listening is also important because it occupies a significant amount of time that L2 educators spend on communicating. Listening also provides input that can be most crucial for L2 acquisition in general or in the development of integrated skills such as speaking and writing. Many of the listening research studies are based on a listening-to-speak or a reading-

listening-writing integration, which is why I use a listening-to-write integration to fill a research gap.

As for **knowledge transfer**, Haskell (2001) defines learning transfer as "use of past learning when learning something new and the application of that learning to both similar and new situations" (Haskell, 2001; Mekala et al., 2016). Knowles (1970) adds that a key component of adult learning is ability to make real-world connections, and he suggests that all learning be transferable to new situations. Haskell supports that transfer is inextricably intertwined with learning. Two factors that Ormrod (2011) provides for improving efficiency of language transfer are: (1) ability to show learning in a meaningful context rather than rote, and (2) ability to show the new information in the correct context.

The TOEFL iBT Test Writing section has an integrated writing task, reading-listening-write where test-takers read passages and listen to lectures, then they respond to questions that ask what they read and heard. More specifically, this integrated writing test involves several steps. First, the test-taker reads a 250-300 word article within three minutes about an academic topic. Second, the test-taker listens to a two-minute lecture about the same topic. Third, the test-taker is given 20 minutes to write an essay about how the lecture challenges the reading. The test-taker can see the article while writing the essay, but not listen to the lecture again. This test exemplifies knowledge transfer. While part of the test is to summarize the lecture, the transfer element is relating the lecture to the reading, which requires test-takers to not only identify main points and select details, but also to connect ideas between the two inputs into a text. The explanation element requires test-takers to write about the contrast of ideas. These skills align with some of the constructs for the TOEFL test that would exemplify transfer, e.g., selecting details, connecting ideas, explaining the way a speaker feels about the content, and drawing inferences/conclusions.

It is important to note that while the definitions of knowledge summary and knowledge transfer have been distinguished, the boundary between the two constructs is not always clear when operationalised in practice. Several scholars (e.g., Keck, 2014; Hirvela & Du, 2013; etc.) suggest that there is an overlap between the two constructs such that the two are not completely separate criteria. Learners often produce language output that fades the boundary between the two definitions. Part of this overlap is that both constructs involve comprehension, organization, and use of input language, i.e., reading, listening, or both. In language production, learner output reflects aspects of both, i.e., summary (restating and condensing information from the input) and transfer (integrating ideas and recontextualizing them into the output). Depending on the demands and purpose of the task, learners often shift between

condensing and restating ideas for the output, such as reproducing key points during some of their practice, while at other times they rephrase and extend ideas from the input that they use for the output. For example, summarizing typically requires rewording and reorganization, skills that are also necessary for transfer (Grabe & Zhang, 2013). Transfer, similarly, often involves summarization skills before learners can apply the information to a new context, thus blurring the line between summary and transfer. This overlap suggests a potential challenge for research, as establishing clear boundaries between the two constructs remains challenging but necessary for clear analysis.

To reflect into my study the definitions of knowledge summary and knowledge transfer discussed in this section, I developed a task that assesses listening comprehension (knowledge summary) and the ability to use new information and make real-world connections to new situations (knowledge transfer). I delve further into detail about this task in the Methodology chapter.

2.8 Chapter Summary – hypotheses and research questions

In sum, the literature review above indicates that a major research gap is the impact that task repetition and feedback have on integrated listening-to-write tasks in relation to the complexity, accuracy, and fluency of students' academic performances, as well as to their knowledge summary and the transfer of information students glean from the listening input and incorporate into writing.

The existence of the relative gap in the literature to investigate the impact on written performances after a series of task repetitions has inspired my pursuit for a listening-to-write integration for EAP in this research. Task skill integration is needed in EAP classes to enhance student engagement with the language, most notably because each language skill rarely functions in the absence of the other skills.

More specifically, my study aims to explore the distinctive effectiveness of repeating the same integrated listening-to-write task at three intervals in relation to CAF, knowledge summary and knowledge transfer. My hypothesis aligns with previous task repetition research such that written performance, like oral performance, improves with repetition. Also, informed by previous research, I hypothesise that during the writing repetition process, some of the attentional foci compete with one another at the first repetition (second performance), and then significant improvements occur at the second repetition (third performance). In addition, I hypothesise that by receiving feedback, it will help students improve their writing. Further, I hypothesise that students will recognise the benefits of task repetition.

The overarching question of this study therefore is: *How does repetition of and feedback on an integrated listening-to-write task impact second language university EAP student writing?* To address the various elements of this overarching question, I raise the following more specific research questions:

RQ 1. Is there a change in listening-to-write task performance scores (CAF, knowledge summary and transfer) as a function of task repetition (Time) and feedback, such that the effect of task repetition (Time) on the task performance scores depends on whether students receive feedback or not?

RQ 2. Is there a change in listening-to-write task performance scores (CAF, knowledge summary and transfer) as a function of task repetition (**Time)**? If so, in which direction?

RQ 3. Is there a change in listening-to-write task performance scores (CAF, knowledge summary and transfer) as a function of receiving feedback or not (**Feedback**)? If so, in which direction?

RQ 4a. What are students' **perceptions** of the task used in this study, of listening-to-write tasks more generally, and of the extent to which integrated task repetition helps EAP students develop their writing proficiency?

RQ 4b. To what extent do student perceptions of task repetition differ between those who received feedback on their writing performances and those who did not?

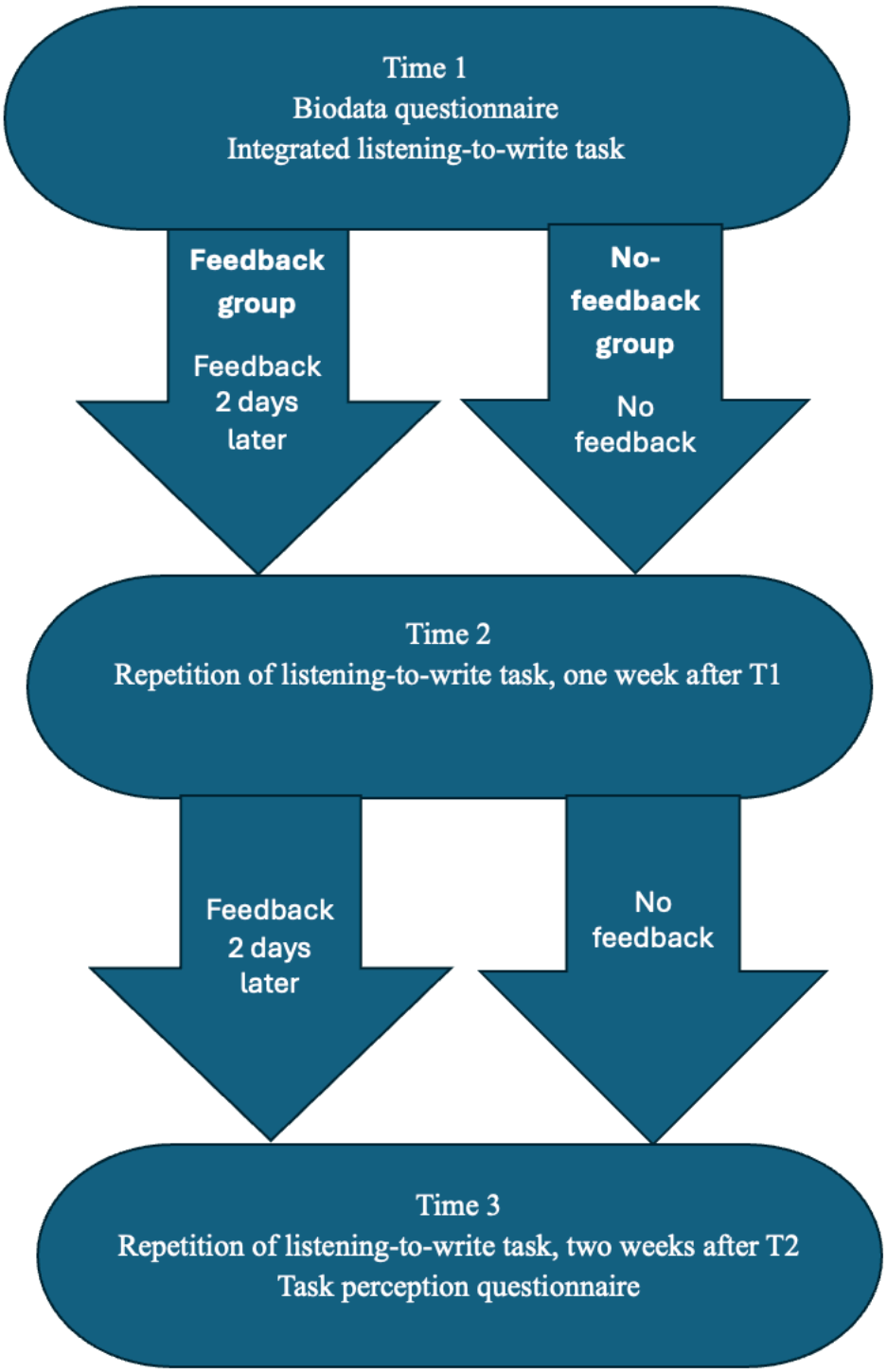
3 Methodology

In this chapter, I first present an overview of the research design (section 3.1). Then, I describe research ethics (section 3.2), the research setting (section 3.3), and the participants (section 3.4). Next, I provide details about the data collection methods and instruments (section 3.5), such as a description of the data collection setting, background questionnaire, listening-to-write task, performance assessment instruments and measures, i.e., CAF measures and rating scales, and the task repetition perception questionnaire. After that, I explain the procedures and feedback methodology (section 3.6). Subsequently, I explain the data analysis methods (section 3.7) for CAF, knowledge summary, knowledge transfer, and task perception questionnaire data.

3.1 Overall research design

To answer the research questions that I presented in section 2.8, I designed a primarily quantitative study that took place at a public university in the United States. In this study, I measured the effects of repetition and feedback on 64 upper-intermediate university EAP students' writing performances in an integrated listening-to-write task that they completed an initial time (Time1), then repeated two times (Time2, Time3). I randomly placed the students into two groups – feedback/no feedback. The feedback group received feedback two days after Time1 and Time2. There was a one-week interval between Time1 and Time2, then a two-week interval between Time2 and Time3. After Time3, I surveyed the students, using a questionnaire, for their perceptions of the integrated listening-to-write task as well as their perceptions of the effectiveness of repeating tasks to improve their writing. Figure 3.1 visualizes the overall research design.

Figure 3.1. *Overall research design*



3.2 Research Ethics

Ethical approval for the study was given by the Faculty of Arts and Social Sciences and Lancaster University Management School Research Ethics Committee (FASS-LUMS REC) at Lancaster University and by the Institutional Review Board (IRB) committee in the Office of Grants Administration of the US university where data collection took place. All participants

were given a participation information sheet (see Appendix 1), then they signed a consent form (see Appendix 2).

To follow Cohen et al. (2018), the informed consent and participant information sheets included explanations and descriptions of the study's topic, purpose, methods, procedures, and of data storage and reporting. In addition, I outlined benefits that might derive from the study, students' right to voluntary non-participation or withdrawal from the study, and "rights and obligations to confidentiality and non-disclosure of the research, participants and outcomes" (Cohen, 2018, p. 125). In compliance with the UK's Economic and Social Research Council's (ESRC) (2015) research ethics principle that emphasizes the importance of respecting anonymity and confidentiality of personal data, I included a statement on the participant information sheet that clarified that all documents would be encrypted, personal information removed, and that pseudonyms would be used in publications and presentations if individual participants' data needed to be referred to.

I recruited the participants through an introduction that their professor made for me during class while I was present. During this introduction, the students had the Participant Information Sheet. The professor and I reiterated to the class, as we had pointed out on the information sheet, that non-participation in my study would not affect their studies or their evaluation in the class. Those who did not want to participate were free to work on their class assignments. None of the students who chose not to participate did the tasks. All of the data collection took place during regular class time in the computer room that was reserved for specific class sessions.

While I do work at the university where this research was conducted, I was teaching in a department other than ESL. The ethical risks were minimized by the fact that the participants were not my own students, and I did not know them. Therefore, I could form an independent evaluation, and there was no potential pressure for the students to participate in the project. I was able to be as independent a researcher as possible because there was no conflict between teacher and researcher roles.

3.3 Research Setting

This study was conducted in an upper-intermediate EAP programme at a public university in a major city in the northeastern United States. The university determines the specific international students (who have already been accepted into their degree programs) who need to be enrolled in the EAP programme based on the results of the initial Accuplacer ESL Test that they first take when they are accepted into the university. The EAP programme runs prior

to degree studies. In this case, the degree study semester starts would be deferred until the students complete their EAP programme.

In the EAP programme at this institution, language immersion students are placed into their EAP class level based on the results of a placement examination consisting of the English Placement Test (EPT) and the Michigan Composition Test (MCT). Students may attend up to three semesters in the EAP immersion programme during the first year of their university studies to develop their English proficiency (see section 1.2 for overall information on the EAP courses' structure and focus). At the end of each EAP course, students take an instructor-led examination to determine whether they passed the course. At the beginning of each semester in the EAP programme, students take a version of the EPT (returning students are given different versions to avoid repetition), based on which the administration determines the EAP course level that the student will take that semester.

The EPT is a multiple-choice test that assesses reading and listening comprehension, vocabulary, and grammar. The MCT is an essay test where students are given 30 minutes to write an essay of 350–500 words on a topic stated on the exam. Two EAP teachers grade the essay portion, and the performances are evaluated on critical response, development and organization, word choice, grammar sentence structure and mechanics (all equally weighted). The EPT counts 40% of the total placement score, and the MCT counts for 60%. After scoring, the university determines the cut scores that are applicable to their student population and class availability, and then places them into one of six levels of EAP classes with similar level students in each group.

It is worth noting that students who do not pass the EAP class(es) are still permitted to pursue their studies. In fact, if they choose to advance to their degree studies without taking any EAP classes, the university allows it. However, these same students do not receive the special funding that they would normally receive if they score higher on the Accuplacer ESL Tests that students take again at the conclusion of their EAP programme participation.

The upper-intermediate level is the second highest level; I explain the reason for focusing on this level in section 3.4.1.

3.4 Participants

3.4.1 Rationale for my selection of upper-intermediate EAP students as participants

Undergraduate students from the EAP immersion programme participated in the study. These students had been invited to take part for four reasons. First, they were a group of EAP learners who needed English for their academic courses. Second, they were pursuing studies in various

majors, thereby representing a wide cross-section of this student population. Third, it was hoped their proficiency level and their experience in using English for academic purposes would assist them in self-assessing whether the linguistic demands of the task type focused on in this study represent L2 learning and use outside of testing conditions (which form the focus of RQ4). This experience in EAP refers to their experience having previously taken lower-level immersion courses, as well as their experience already having started taking an upper-intermediate-level EAP course. Fourth, these students were in the same level English course, thus I could control for proficiency to create parallel groups for feedback conditions.

A key reason for conducting the study with upper-intermediate students was that the integrated task that was the basis for this research was designed for higher-level language learners and these would be more likely to be able to succeed at completing the task. Because the writing task that I used for this study required the application of several language skills, I needed participants who could express themselves through writing quite intelligibly and accurately while using integrated listening and writing skills. Like writing, listening is an important skill in English language learning, as also discussed in section 2.7, especially in university contexts such as during lectures. Without this skill, students are not capturing some of the nuances found within the flow of the spoken language such as recognizing vocabulary through sounds, understanding the lecture material, comprehending lengthy input, understanding relations between different parts of input, keeping track of main points and distinguishing them from details, inferring, acquiring both formal and colloquial language, etc. Also, listening requires people to absorb the input more actively because, unlike reading, listening input such as that from a conversation or lecture is not normally supplemented with a transcript for the listener to refer back to for details (although technological advancements are starting to change this, but this was not common yet at the time of my study). Further, these are among many of the key skills that language programs including the one at this university instill into the daily curriculum. Therefore, I was conducting this research using skills that aligned with the department and particularly the higher-level courses.

According to the Council of Europe (2023), a B2 qualification – which is often labelled as upper-intermediate – suggests that a learner can "understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation…[and] produce clear, detailed text on a wide range of subjects and explain a viewpoint…." (para. 4). Fleckenstein et al. (2020), compare criteria in a TOEFL integrated essay rubric to those of CEFR students at a similar level. For example, written responses that effectively address the topic and task, are very well organized and developed while using

appropriate explanations, along with demonstrating syntactic variety with only minor lexical or grammatical errors are well structures, address argumentative positions, and "convey information in a unified and coherent way…[as well as] express themselves correctly in writing" (p. 4). These are also skills needed to complete the task in my study, thus another reason why my task would be most suitable for CEFR B2 level students (and upwards). Also, based on my experience having taught previous cohorts of students, upper-intermediate learners at my institution hold a good range of vocabulary, have a good understanding of grammar and syntax, and practice effective fluency such that they can make sense of the communicative context without direct studying or learning, thus are able to express themselves on a variety of topics even though they may use target language that is not always correct.

Not only is it a common assumption at the university that the upper-intermediate level is suggested to be equivalent to B2, but Michigan Language Assessment (2020) describes how the Michigan Test is aligned with CEFR. In terms of the B2 level, a student is able to (1) "understand the main idea of complex texts on both concrete and abstract topics, including technical discussions in his/her field of specialization"; (2) engage with highly proficient English language speakers without strain; and (3) "produce clear, detailed texts on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options" (p. 1).

Finally, while students from an advanced level would also have been appropriate participants, there was a larger potential participant pool in the upper-intermediate level at my institution.

3.4.2 Participant profile

Sixty-four students agreed to take part in the study. Thirty-five were male and 29 female. Their ages ranged from 17 – 28 years old (M=22.05, SD=3.119). They were registered to enter the following degree programmes upon successfully completing their EAP courses. Areas of study comprised of Humanities (n=10) Business (n=10), Legal (n=3), Technology (n=6), Education (n=3), Environmental Sciences (n=1), Health Professions (n=12), Mathematics (n=4), Physical Sciences (n=1), Social Sciences (n=2), Art (n=3) and Other/Undeclared (n=9).

All participants were international students who had had 1 – 4 years of English language study (M=2.27, SD=0.996) and spent 1 – 5 years in an English-speaking country (M=2.25, SD=0.992). Their L1s were Spanish (n=22), Arabic (n=4), Italian (n=1), Portuguese (n=1), Chinese (n=8), Korean (n=5), French (n=8), German (n=1), Polish (n=2), Wolof (n=2), Albanian (n=2), Swedish (n=1), Japanese (n=1), Macedonian (n=2), Turkish (n=1), Farsi

(n=1), and Russian (n=2). They were from three different upper-intermediate EAP classes of the programme, and I randomly placed the students evenly into two groups: Feedback group (32) and No-Feedback group (32).

3.5 Data Collection Methods and Instruments

3.5.1 Biodata questionnaire

To establish a profile of the participants, biodata were collected through an online personal background questionnaire administered with the software Qualtrics (see Appendix 3) before the beginning of the task at Time1. The biodata questionnaire contained questions about participants' age, gender, L1, nationality, number of years of speaking English, area of study, and self-ratings on their various language skills using a 5-point scale (Beginner = 1; Basic = 2; Intermediate = 3; Upper-Intermediate = 4; Advanced = 5). I piloted the biodata questionnaire to a small group before I conducted it in the main study where I replicated its administration in the same manner. No revisions were required based on my pilot.

3.5.2 Listening-to-write task

3.5.2.1 The task

The task I developed for this study was an integrated listening-to-write task, which I adapted from materials on Randall's ESL Cyber Listening Lab (www.esl-lab.com) Further, I piloted the listening-to-write task to a small group before I conducted it in the main study where I replicated its administration in the same manner. No revisions were required based on my pilot. Figure 3.2 shows this task prompt.

Figure 3.2. *Listening-to write task prompt*

Listening-to-Write Task: Success in Your Class

'You will now listen to a professor discuss important things that students must know on the first day of class. When the recording has finished, you will be asked to write a short essay about the way you plan your university classes.

Imagine you are a student in this professor's class. A number of factors may play a role in drawing up your university class schedule, including some of the following things you will hear.

Discuss your answers to the following questions. Explain your answers. Do not separate your answers. You have 30 minutes to complete this writing task.

- What are three (3) things the professor talked about in the recording?
- Tell about at least three (3) things you think would be good to do from Day 1 onward to succeed in the class. Discuss the benefits of each of the things you do.
- Discuss at least three (3) qualities you look for in an excellent professor. Explain why.

The website where this task was adapted from provides listening recordings and quizzes for academic purposes to prepare for tests such as TOEFL. Given that the study participants were preparing for degree-level study, they were similar to TOEFL preparation populations as this test is required for university entry by many institutions around the world, including the United States. Thus, the ESL-lab site materials were considered appropriate for the participants.

The listening input I selected for this study was a one-minute recording of the first day of an Intercultural Communications course when the professor introduces the course by describing some key information from the syllabus including the schedule, room, textbook, grading, office hours and lab time. (see Appendix 4 for a transcript). The ESL-lab website for this listening input also includes some activities ('quiz'). The first of these are multiple-choice questions requiring learners to select the correct answers based on what they heard in the listening passage. However, to better represent the academic language domain and to assess the students' abilities to develop their own summaries and draw conclusions from the lecture, I did not use the pre-existing multiple-choice questions such as "The class meets from___" where one selects from the choices the correct time frame. Instead, I created a listening-to-write task for the listening input by adding the writing prompt. To make the task more purposeful, I created some context in the introductory part of the task instructions (see Figure 3.2). A key

characteristic of the purpose-recorded listening input is it includes authentic features such as a room full of students chatting before the lecturer starts talking and hesitations such as "uh" by the lecturer. The motivation for my choice of the input is that it is broad enough to be accessible to students from a wide range of fields, i.e., something that is likely to be covered by lecturers across all fields. For example, there is no technical subject knowledge required to understand the input. For the actual writing instructions, I created three parts to the prompt ('questions'). I provide my rationale for the three elements in the following section.

3.5.2.2 Rationale for the task

Many standardized tests, including the TOEFL iBT, include tasks that require test takers to "synthesize information across listening and reading components of the prompt and then respond in writing, adding another skill to consider" (Plakans, 2016, p. 162). Unlike an independent task that assesses one skill, integrated tasks represent a more challenging construct regarding skill elicitation and content integration. As discussed in the literature review, integrated skill use is also a characteristic of academic language use. Thus, I opted for an integrated task – more specifically, I created the integrated listening-to-write task shown in Figure 3.2. This writing task that is based on a recorded lecture reflects the type of some of the tasks used in the EAP programme where this study took place.

With regard to the writing instructions, I developed the following first question: "What are three (3) things the professor talked about in the recording?" (Question 1) to elicit responses that retell information from the listening, so *knowledge summary*. Both restating facts and summarizing information also align with the descriptor "selecting the important information from the lecture and coherently presents this information…." on the TOEFL rating scale used in this study (see Appendix 5 and section 3.5.3.2). Next, I changed the original question from ESL-lab "What are the key features of the class" to "Tell about at least three (3) things you think would be good to do from Day 1 onward to succeed in the class. Discuss the benefits of each of the things you do" (Question 2). Question 2 was about understanding the key features of the lecturer's course description and then transforming these into action points, so a very direct form of *knowledge transfer*. My reason for this change was to provide more context, make it more purposeful and authentic, and essentially to make it more of a task. The final question that I developed was "Discuss at least three (3) qualities you look for in an excellent professor. Explain why" (Question 3). With this question, it required the student to reflect on this input as a whole and then connect this to what they think are good qualities for a professor. In this case, it started off from overall comprehension of the input, and then

transferring that to a further academic topic that is related but not literally covered by the input. Thus, to stimulate transfer, I used the words “Discuss” and “Explain” in the prompts to require supporting details and argumentative writing. So, this is a broader and more distant *knowledge transfer* than Question 2.

Various abilities that a B2 level student needs also relate to my task. According to the Council of Europe (2023), a CEFR B2 qualification – which is often labelled as upper-intermediate – suggests that a learner can “understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation…[and] produce clear, detailed text on a wide range of subjects and explain a viewpoint….” (para. 4). According to the Council of Europe (2020), explaining a viewpoint also includes the ability to “develop an argument giving reasons in support of or against a particular point of view” (p. 174). Going a step further, B2 level students can “use a variety of linking expressions efficiently to mark clearly the relationships between ideas” (p. 174). In terms of oral comprehension, an example of an ability at the B2 level is “understand standard language or a familiar variety, live or broadcast, on both familiar and unfamiliar topics normally encountered in personal, social, academic or vocational life” (p. 48). In terms of overall writing production, an example of an ability at the B2 level is “produce clear, detailed texts on a variety of subjects related to their field of interest, synthesising and evaluating information and arguments from a number of sources” (p. 66).

Fleckenstein et al. (2020) compare criteria in a TOEFL integrated essay rubric to those of CEFR students at a similar level. For example, written responses that effectively address the topic and task are very well organized and developed while using appropriate explanations, along with demonstrating syntactic variety with only minor lexical or grammatical errors. They are well-structured, address argumentative positions, and “convey information in a unified and coherent way…[as well as] express themselves correctly in writing” (p. 4).

The redeveloped questions in my task require more extensive writing which enables raters to observe students’ abilities in knowledge transformation, as, for example, defined in the “Transfer” descriptor on the Association of American Colleges and Universities (AACU) rating scale, which is used in many universities across the United States for similar academic contexts (see Appendix 6 and section 3.5.3.2). These all assess the application of information use from one situation to new situations, as prescribed by AACU.

3.5.2.3 This task as conceptualized in TBLT

This task exemplifies task definitions as conceptualized in much of the literature described in the previous chapter. For example, its primary focus and outcome are on meaning, a key pillar in researchers' categorizations of a task (e.g. Nunan, 1989; Willis, 1996; Ellis, 2003, 2009; Skehan, 1989; Prabhu, 1987). The outcomes of this task are non-language driven (Richards & Rodgers, 2001), a key criterion of researchers' explanations of tasks (Ellis, 2003; Skehan, 1989). Namely, students use their own linguistic resources in the sense that by writing about what they must do to succeed in college (Q2) and things they seek in an excellent professor (Q3), they must focus on content and come up with their own ideas. They are provided with a time limit to have some sort of practical control, but there is no word list they must use nor a word limit. Therefore, they can focus on completing the content rather than potentially being distracted by counting the words they produce. Similarly, no linguistic criteria are explicitly listed in the prompt. In the integrated task, there is still a linguistic outcome in the sense that students must explain in written form about the way they will apply the information that they learned from the input. However, the main goal is to convey meaning.

Table 3-1 summarizes how this listening-to-write task aligns with the common criteria that reflect the definition of a task collectively defined by researchers.

Table 3-1. *TBLT criteria and application in this task.*

TBLT criterion	Application in this task
Primary focus is a meaningful activity	This task is based on information provided in the recorded lecture. This is an academic lecture excerpt, reflecting what might be an initial session at the start of the course.
Dependency on communication and interaction skills	Use of own linguistic skills without a script or word list
Gap where students must convey information	Students are retelling information that they recalled from a lecture as if they were conveying it to someone who missed the lecture (Question 1: What are three (3) things the professor talked about in the recording?)
Learners should need to rely on their own linguistic or non-linguistic knowledge	This task is not based on the use of any specific linguistic form. Students need to come up with their own additional content and language skills because there is content in the lecture, thus more is needed (Question 2: Tell about at least three (3) things you think would be good to do from Day 1 onward to succeed in the class. Discuss the benefits of each of the things you do; and Question 3: Discuss at least three (3) qualities you look for in an excellent professor. Explain why.)
Target language is used in a meaningful way to finish the task as opposed to producing specific forms	Students are applying information they learned in the lecture by explaining the way they could use the information and apply in other settings such as in their university. No specific grammar focus is required to be used (Questions 1-3)

In line with the TBLT literature, the information gap is where students identify from the listening passage on the orientation day in class some key items that the professor discussed as they, for example, would email to a classmate who missed class. More precisely, the first question of the task asks to retell three things that the professor talked about, and thus aligns with part of Ellis's (2003, 2009) task meaning that requires learners to convey information. The second question asks learners to tell about things they think would be good to do and their benefits, and thus aligns with Ellis's (2003) and Prabhu's (1987) reasoning gap where learners use existing information to draw inferences to create new information. The third question asks to explain qualities they look for in excellent professors, which aligns with expressing an opinion as emphasized by Ellis (2003, 2009) and Prabhu (1987). Also, Willis's (1996) inclusion of problem-solving aligns, admittedly somewhat more weakly, with the second and third question that serve as prevention of potential academic problems. In the second and third question, students select ideas that they gleaned from the recording that they then identified as recommendations that they would make to succeed in class as well as what they look for in an excellent professor. This aligns with an opinion gap. It also addresses part of Willis's (1996) meaning to include listing.

All three questions reflect Ellis's (2009) outcome requirements, i.e. other than solely target language, and Richard and Rodger's (2001) requirement of non-language-driven communication outcomes that also show relevance to learner needs such as student success in college. The questions also reflect some of Skehan's (1989) task characteristics including the relationship to real-life activities such as summarizing and drawing inferences from missing class notes.

Révész's (2011) commentary about opportunities for linguistic error feedback on meaning-based tasks is my rationale for exploring the role of feedback through the experimental design of this study. Skehan's (1989) commentary on importance of language use for task performance and for communicative purpose justifies my use of separate assessment measures for fluency and knowledge summary and transfer, which I describe in the next section.

3.5.3 Performance assessment instruments and measures

To be able to answer RQ1, RQ2, and RQ3, students' written performances on the listening-to-write tasks were assessed in two different manners: by means of CAF measures and rating scales.

3.5.3.1 CAF Measures

Here, I describe and justify which CAF measures I selected to analyse linguistic aspects of students' written performances on the listening-to-write tasks.

Complexity

Table 3-2 lists the measures used to establish the complexity of the listening-to-write performances.

Table 3-2. *Complexity measures*

Construct	Measure	Evidence of improvement	Analysis
Sentential complexity	Average sentence length	Increase in number of words per sentence	MS Word & Manual analysis
	Ratio of simple sentences to complex sentences	Decrease of simple sentences in proportion to complex sentences	Manual analysis
	Clauses per T-unit (C/T)	Increased number of clauses in proportion to T-units	Manual analysis
Lexical complexity			
Lexical diversity	MATTR50	Higher range of different words within a text	TAALED webtool
Lexical sophistication	Sophisticated words & lexical sophistication	Increased use of sophisticated words & increased number of sophisticated lexical words in proportion to total lexical words	Lextutor lexical analysis webtool

Complexity in writing was defined in this study as grammatical and lexical complexity. As learners become more proficient in a language, their writing becomes more complex, and they use more sophisticated vocabulary and sentence structures. For example, Martinez (2018) states that as L2 learners progress, they may write fewer short simple sentences but connect their ideas using compound and complex sentences, thereby increasing dependent and independent clauses, and writing longer sentences with more ideas.

The measures I adopted for grammatical complexity and, more specifically sentential complexity, are *average sentence length*, *ratio of simple sentences to complex sentences,* and *clauses per T-unit [C/T]*. Norris and Ortega (2009) consider the average sentence length an indicator of complexity rather than fluency. I manually counted the number of sentences, simple sentences, compound sentences, complex sentences, dependent clauses, and independent clauses, then recorded the count for each text to compare among task completions.

An increase in average sentence length and/or C/T would be an increase or improvement in complexity; an increase in the ratio of simple sentences to complex sentences would be a decrease in complexity. A T-unit, as introduced earlier, is a sentence or a part of a sentence with a main clause, along with any attached clauses or non-clausal structure. Gass and Selinker (2001) define a T-unit is a sentence or a part of a sentence that can stand as a grammatically complete sentence on its own if it is punctuated like a full sentence.

Previous researchers consider more writing as signs of complexity to include elongated sentences, subordinated structures, and longer T-units with one or more subordinated structures (Hunt, 1965; Lintunen & Makila, 2014; Bulté and Housen, 2012). Holger (2004) adds that "complex sentences originate from simple sentences that are gradually linked together, through coordination and subordination. This linking of production units makes the language more complex" (p. 3). For this reason, I counted compound sentences as part of complexity when I calculated the ratio of simple sentences to complex sentences. A simple sentence is a complete sentence with only one independent clause. A compound sentence is one that has two independent clauses that are joined together by a coordinating conjunction such as "and," "but," or other coordinating conjunctions. A complex sentence has at least one independent clause and a dependent clause.

This ratio of simple sentences to complex sentences as another sentential complexity measure was also used in Davison's (2021) study. I further calculated sentential complexity (i.e., linguistic competence) by dividing the number of clauses by the number of T-units *(C/T),* as used in previous research (Mady, 2018; Wang & Jin, 2022; Kyle, 2021). I considered whether these advanced structures were appropriate for upper-intermediate level use for this study.

For lexical complexity, I focused on lexical diversity and lexical sophistication as they have been described as the main distinctive aspects of lexical use (Michel, 2017). Lexical diversity, as introduced in earlier, refers to "the range of different words used in a text, with a greater range indicating a higher diversity" (McCarthy and Jarvis, 2010, p. 381). In the latest decade, researchers typically used the lexical diversity measure D, but improved measures now exist based on the latest research by Kyle (2023). Consequently, I measured lexical diversity using MATTR50 in TAALED (Kyle et al., 2021). If a repeated text has a higher MATTR50 score than an earlier text, the writer used more variety of words. Lexical sophistication is the "percentage of sophisticated or advanced words in a text" (Lindqvist et al., 2013, p. 110). I used the Lextutor lexical analysis webtool to count the number of sophisticated words (Cobb, 2002; Heatley et al., 2002). I further calculated lexical sophistication by dividing the number of

sophisticated words by the total number of lexical words. If lexical sophistication increased, then there was a higher number of sophisticated words in proportion to total lexical words.

Accuracy

Table 3-3 lists the measures I used to establish the accuracy of the listening-to-write performances.

Table 3-3. *Accuracy measures*

Accuracy			
Construct	Measure	Evidence of improvement	Analysis
Grammar: • Subject-verb agreement • Verb tense • Verb form usage • Prepositions • Articles	Errors per 100 words	Decreased number of errors per 100 words	Manual analysis
	Errors per T-unit (E/T)	Decreased average number of errors per T-unit	Manual analysis
	Error-free T-units per T-unit (EFT/T)	Increased average number of error-free T-units per T-unit	Manual analysis

In section 2.3.2, I introduced some accuracy measures that have been used in previous studies. I also discussed some commonalities of the types of errors used in this study. As Plakans et al. (2016) observed, "Accuracy in writing has been captured by counting errors, calculating ratios of phrases, clauses, or T-units with and without errors, using holistic/analytic rating, and weighted error ratios" (p. 164). Pallotti (2019) suggests evaluators use global measures to measure accuracy, notably number of errors per unit or per error-free unit. This aligns with recommendations by Wolfe-Quintero, Inagaki, and Kim (1998) for accuracy measures: *errors per T-unit (E/T)* and *effort-free T-unit per total T-units (EFT/T)*. Thus, I adopted these measures in my study. Pallotti (2020) states "It remains difficult to define exactly what errors are, especially if lexical, morphological, syntactic, phonological or spelling errors are bundled together" (p. 203). Therefore, I separated my specific measures in addition to bundling them together. Additionally, because the number of words that the students wrote at each time in this study were not the same, to control for those differences in text length, I also calculated *errors per 100 words*. An increase in errors per 100 words and/or E/T would be a decrease in accuracy; an increase in EFT/T would be an increase or improvement in accuracy.The linguistic targets I measured were: subject-verb agreement, verb tense, verb form, articles, and prepositions. Earlier, I discussed some of the commonalities of these types of errors. Before I selected these linguistic features, I informally interviewed two university EAP instructors at my institution and they also showed me a total of 41 recent student essays so I could identify common errors. Based on these conversations and reviews of student

writing, the error types listed above were the most common errors made by their students in their regular coursework, thus my decision to focus on these in this study. Like other researchers (e.g. Davison, 202; Kormos, 2014; Kuiken & Vedder, 2012), the grammatical errors in my study were counted separately (by error type) before also being counted together to constitute a total number of errors (global). I manually counted the errors in all writing samples. Then I used the Sentence Extractor + T-Unit Calculator tool to count the T-units (Cobb, 2017). Afterwards, I manually counted the T-units to check for errors. A second rater assisted in marking a subset of my texts, and more details about inter-rater reliability for this study are provided in section 3.5.3.2.1.It is important to note that studies that look at global measures versus specific linguistic grammatical measures might report differing findings. Ortega (1999) argued that by using both kinds of measures, it would provide a more focused overview of accuracy in a study than if only global measures were used. In fact, a study that used only global measures will "have the disadvantage of being too broad to capture small changes…and [will] obscure errors in grammatical domains that may be important at a given level of development" (Ortega, 1999, p. 118). In the Results chapter, I will provide the findings for accuracy by global measure (see section 4.1.2) and by error type (see section 4.1.3).

Fluency

Table 3-4 lists the measures I used to establish the fluency of the listening-to-write performances.

Table 3-4. *Fluency measures*

Construct	Measure	Evidence of improvement	Analysis
Written language proficiency	Words per T-unit (W/T)	Increased average number of words per T-unit	Manual analysis
	Words per error-free T-unit (W/EFT)	Increased average number of words per error-free T-unit	Manual analysis

As introduced earlier, fluency refers to the rapidity that learners can produce language within a prescribed temporal period (Skehan, 1998a). In writing, fluency is typically measured by length of the text produced and errors in the text. I used measures that reflect these two aspects and which many previous researchers have used (Ghahderijani, 2021; Barrot & Agdeppa, 2021; Davison, 2021; Wolfe-Quintero, Inagaki and Kim, 1998; Gass and Selinker, 2001): *words per T-unit* (W/T) and *words per error-free T-unit* (W/EFT). These are considered effective fluency ratio measures because they measure the student's ability to write

longer and more advanced sentences with improved fluency in writing. While my task had no word limit, there was a 30-minute time limit. An increase in W/T and/or W/EFT would be an increase in fluency.

Many studies confirmed T-units as a reliable measure of syntactic development because they are easy to identify and measure (Wolfe-Quintero, Inagaki and Kim, 1998; Gass & Selinker, 2001; Mackey & Gass, 2005). Various studies have shown that it is a reliable indicator of syntactical development (Larsen-Freeman, 1983; Bardovi-Harlig & Bofman, 1989). Because T-unit count measures alone do not suffice for syntactic development, error-free T-unit analysis is an additional measure used to assess the number of errors in relation to the sentence length. This helps confirm whether writing length increased with improved linguistic accuracy. This is a reason why I include this as a measure.

The numbers of words, T-units, and errors were identified as previously stated. To calculate the W/EFT, I first manually counted the error-free T-units, then I recounted the number of words in each writing sample with error-free T-units only with the T-units with errors excluded from the equation. To calculate W/T, I divided the total number of words by the number of T-units in the text.

In sum, I presented the CAF measures employed in this study and their ground for selection. In principle, more and other CAF measures are available for analyzing written performances. However, "CAF analytic measures are unlikely to be directly used to arrive at a test score, both because calculating them is time-consuming and…because human judgement is often necessary to determine the communicative value of these linguistic aspects" (Pallotti, 2020, p. 207). This is one reason why I did not select more CAF measures, i.e. linking this back to the university context for my study, is that they are not thoroughly relevant to communicative approaches to language teaching and testing. Therefore, I investigated only a small selection of, for example, accuracy and fluency measures, and not more. Pallotti (2020) concludes by stating "In practice, many rating scales and test scores will be based on a number of [CAF measures] at the same time, but this should be the result of an intended choice, with a clear awareness of what (sub)dimensions have been bundled together any way" (p. 206).

3.5.3.2 Rating scales

Table 3-5 lists the measures I used to establish *knowledge summary* and *knowledge transfer* of the listening-to-write performances.

Table 3-5. *Knowledge summary and transfer measures*

Construct	Measure	Evidence of improvement	Analysis

Knowledge summary	TOEFL integrated task rubric (modified)	Increased holistic score	Manual analysis
Knowledge transfer	AACU rubric (modified)	Increased holistic score	Manual analysis

Two existing rating scales were adapted for the purpose of this study to measure *knowledge summary* and *knowledge transfer*. The first rating scale that I adapted to measure student performances in writing (specifically, the construct of *knowledge summary*) was the Integrated Writing Rubric (Scoring Standards) developed for the TOEFL Test (Test of English as a Foreign Language) (see Appendix 5). This scale was felt to be particularly suitable because it is designed specifically for an integrated task, for academic purposes testing, and for a test-taker population like the participants in this study. It has also been validated by other researchers for use in large-scale integrated testing (Plakans, 2009; Gebril, 2006; Gebril & Plakans, 2008). The scale contains six rating bands that provide performance descriptions for each score. Although it includes descriptors for errors of usage and grammar, accuracy was measured separately as part of the CAF analyses in this study (see 3.5.3.1), so the scale was amended to exclude such descriptors. The scale includes descriptors for knowledge summary, reflecting test takers' use of information from the lecture and success at presenting this information in their writing. This criterion was retained for the purposes of the present study.

A second rating scale (see Appendix 6) developed by the Association of American Colleges and Universities (AACU), was also drawn upon to rate *knowledge transfer* in this study, a criterion widely used in the US academic context but not included in the TOEFL scale. Knowledge transfer concerns the ability to apply skills, theories, and methodologies from one situation into new situations. This additional criterion is necessary to reflect students' connections of new information to real-life situations, as introduced and defined in section 2.7 where I discussed the connections between the input (listening) with the writing. This is particularly relevant to this study's research setting since US colleges must develop learning outcomes that require use of new materials through interactive liberal learning education, which is also a significantly increasing skill set required by employers according to a 2013 Hart Research survey (Robbins, 2011; Baker, 2009; AACU, 2013).

3.5.3.2.1 Inter-rater reliability

As an experienced EAP lecturer, I have 15+ years' experience rating students' academic writing using rating scales. In my study, I acted as the first rater to use the two holistic rating scales for knowledge summary and knowledge transfer. In addition, I acted as the first rater to

count the accuracy errors listed on Table 3-3. Further, I acted as the first rater to count the T-units and the types of sentences, i.e., simple, compound, and complex.

To ensure rating reliability, I assigned a second rater, another university EAP educator with 10+ years of experience, to rate 25% of randomly selected performances (16 of the 64 students x three performances each (Time 1, Time 2, Time 3) = 48 texts) using the same rating criteria. Mackey and Gass (2005) state: "It is possible to establish confidence in rater reliability with as little as 10% of the data" (p. 243). To avoid familiarity bias, the students' names were encrypted with student numbers to de-identify the texts (and note that the students had typed the texts, so there was also no risk of handwriting recognition).

The rating scales for knowledge summary and knowledge transfer measures, as introduced in section 3.5.3.2, are holistic ones. Therefore, it is important to gauge the consistency in marking scripts because subjectivity in marking can be a potential risk to inconsistent scoring, i.e., subjectivity in scoring might result in different scores by two or more raters for the same text. A high inter-rater reliability score is an indication that the scoring is more likely to be replicable by future markers.

After familiarizing the second rater with the task, rating scales, and other scoring criteria, i.e. CAF measures, we first independently scored two texts to practice using the two holistic scales as well as counting the errors, T-units, and types of sentences. Then, we compared the scores on these first two sample texts. Where variations of scoring occurred, i.e., knowledge summary scores and manual counts for T-units, we conferred by reviewing and discussing the differences between the scoring bands for knowledge summary. We also reviewed T-units. The main reason for the initial differences in ratings were that the second marker counted grammatical errors toward the knowledge summary score. I reminded the second rater that the grammatical errors are counted manually for the accuracy measures only and scored separately from knowledge summary for this study, i.e., the "occasional language errors" and "errors of usage and/or grammar" aspects of this scale were removed for this construct. Next, we rated another two texts, after which time we were able to resolve our differences between scorings, i.e. the inter-rater reliability by simple percent calculation was 100%. Because we both reached high inter-rater reliability between us for those two texts, the second rater and I then proceeded to independently rate the remainder of the 25% of texts for double rating.

Inter-rater reliability was assessed by Cohen's Kappa (κ). We reached moderate to strong levels of agreement on the texts (n=48; McHugh, 2012), with the inter-rater reliability at $\kappa = .79$ for error counting, $\kappa = .799$ for T-unit counting, $\kappa = .887$ for type of sentence counting,

$\kappa = .867$ for type of clause counting, $\kappa = .824$ for knowledge summary scoring, and $\kappa = .759$ for knowledge transfer scoring. Given the satisfactory results, the remaining 144 texts (75% of the performances) were marked by just myself.

3.5.4 Performance feedback

As shown in Figure 3.1, two days after completing the task performance at Time1 and Time2, I provided the feedback group with two forms of focused feedback. In terms of knowledge summary and transfer, they received holistic scores from the adapted TOEFL and AACU rating scales (Appendix 5 and Appendix 6, respectively). These students were also provided with a copy of the two scales, to read the descriptor corresponding with their score. In terms of CAF, they received coded metalinguistic feedback on their linguistic performance based on the error code correction symbol sheet (a standardized sheet used in regular EAP work provided by the university where the participating students attended) that can be found in Appendix 8. To view the features of the accuracy measures that I focused on in the metalinguistic feedback, along with evidence of improvement and analyses, see Table 3-3. The no-feedback group did not receive any feedback, scores, or coaching. Because the participants were in the same class, there was a risk that the two groups might share feedback. However, the groups were instructed to not communicate with one another about any aspect of the task or research.

It is important to note that various factors were taken into consideration when I determined which forms of feedback to provide to the feedback group. There is some controversy about the way errors should be corrected as well as the frequency in which they should be calculated (e.g., Ferris, 2010; Bitchener 2008; Narmina, 2024; Lee, 2013). Also, there are arguments in favor of feedback that is focused (e.g., Storch & Wigglesworth, 2010; Ferris, 2002; Lee, 2019; McMartin, 2014).

A key underpinning in support of focused feedback similar to what I incorporated in my study is to provide feedback that not only focuses on specific types of errors but also feedback that is practical for the students. For example, focusing on fewer error types helps students feel less overwhelmed and, thus, would ideally be more motivated to improve. Lee (2013) stated that "focused written corrective feedback is also better for students, as their papers are no longer inundated with red ink, which is likely to hurt their ego and damage their confidence in writing" (p. 109). Earlier researchers, e.g. Bitchener (2008) and Ferris (2010), also support that focused feedback is a means to help students improve their written performances. At the same time, Bitchener (2008) and Ferris (2010) recommended a common middle ground such that several error types be used for feedback. Bitchener (2008) suggested that "some researchers feel that selecting only one error type for feedback is not practical" (p.

108), therefore suggesting that several error types but not an excessive number be used in the feedback. In fact, Lee (2019) emphasized the importance of "focused written corrective feedback, i.e., responding to errors in a selective, focused manner…[less is more]…." (p. 524).

Researchers and educators need to determine which errors to focus on for feedback such as what I had to determine in my own study. In terms of selective feedback, Narmina (2024) emphasized on the importance of providing feedback for grammatical errors: "Grammar is a fundamental part of language. Without a sound knowledge of grammar, we cannot express our feelings, thoughts, and ideas accurately" (p. 102). At the same time, the students' needs must also be considered. Narmina (2024), for example, indicated "the right approach is to correct only for a helping purpose" (p. 100). Bitchener and Ferris (2012), like Narmina, emphasized the importance of factoring in students' needs. Going a step further, Lee (2013) emphasized the need to recognise student levels: "Teachers have to factor students' needs and proficiency levels into their decision-making" (p. 114). Thus, in terms of the grammar structures that I selected for my study, I factored into consideration the needs of the upper-intermediate level students. As mentioned earlier, I conferred with regular EAP staff at the institution to identify some of the most common errors made by the students.

Motivation plays a role in feedback effectiveness. Next are some studies where my focus on feedback on grammatical aspects of writing aligns with theirs but, in some cases, where there is some student appreciation of feedback and willingness to improve in their grammar. Results from a study conducted by Benson and DeKeyser (2019) on written corrective feedback on verb tense accuracy showed that language learners who received metalinguistic feedback had significantly reduced their simple past tense error rate from 20% to 12%. This positive outcome of receiving feedback is similar to participants from Diab's (2015) study for pronoun agreement errors, Shintani and Ellis's (2013) study for indefinite article errors, and Shintani et al.'s (2014) study for conditional verb tense, "all of whom found that metalinguistic feedback on its own resulted in significant gains in accuracy at the time of the immediate posttest…but lost by the time of the delayed posttest" (p. 717). Building on metalinguistic feedback, results from a survey conducted by Carparas (2020) who reviewed pre- and post-tests of writing samples from 40 ESL students revealed that "error in grammar (82%) is the top most feature the students would like their teacher to correct in their writing" (p. 682), thus a positive perception students have on error correction. Further, "most of the participants strongly agreed that the use of metalinguistic corrective feedback with the use of correction symbols facilitate their revision tasks" (p. 682).

Proper delivery of the feedback is another important factor to consider for it to be effective for this student population. Timeliness is one such factor. Lee (2013) stated "there should not be an unproductive time gap between when the error was made and when it is corrected" (p. 114). The EAP students at the university where I conducted my study typically receive feedback on essays from their professors within a few days. My decision to design my study to follow a similar pattern, i.e., two days after the written performance, aligns with Lee's (2013) point about a productive time gap between performance and feedback. In addition, it aligns with Lee's (2019) point about the "less is more" approach. Frequency of feedback delivery is another factor, notably as it is not ideal for students to become less motivated and more frustrated by receiving excessive amounts of feedback beyond what is appropriate for the specific student populations.

3.5.5 Task repetition perception questionnaire

To be able to answer RQ4a and RQ4b, student perceptions of integrated task repetition were assessed by means of responses to a post-task repetition perception questionnaire. It was administered online in Qualtrics (see Appendix 7).

My method of designing and operationalizing this questionnaire followed Cohen et al. (2018). When I started to design this questionnaire, I first identified the general purpose (to gather student perceptions of task repetition). I developed questionnaire statements into "concrete, researchable fields about which actual data can be gathered" (p. 472), that captured student perceptions both about this task and about repeating tasks in general in relation to developing their L2 proficiency. Following Rukthong (2015), I asked a separate set of questions about the task used in this study because "this method seemed usefully able to investigate the relationship between task perceptions and performances" (p. 99). I then set up two areas of statements where the students could tick their level of agreement on a five-point Likert-scale. Two areas of foci in my scale probed into performance and task motivation, of which these foci were also present in a test-taking motivation questionnaire in Kormos et al.'s (2020) study. My scale follows Rukthong's (2015) questionnaire regarding the point in time of its administration after the last task repetition was completed: "questionnaires, particularly if they use closed-response items, can provide quantitative data (versus qualitative data)…it was preferable to collect the data immediately after task completion in order to obtain as accurate as possible data" (Rukthong, 2015, p. 99). In fact, in my study, the questionnaire was administered just as students had completed their Time3 integrated writing task instead of after each task repetition; thus, to keep the demands reasonable for the students, I kept the questionnaire relatively short. Further, I piloted the questionnaire to a small group before I

conducted it in the main study where I replicated its administration in the same manner. No revisions were made based on the pilot.

In terms of rationale for establishing participants' perceptions, validity and level of difficulty are key pillars of developing an assessment or activity. A valid test or assessment is one that accurately assesses what it is supposed to measure. Bachman (2005) further defines a key characteristic of a valid test as one which test takers perceive as relevant to real-world connections, thereby representing "test appeal" and close alignment to authentic situations. Weir (2005) adds that it is important to read test takers' commentaries about their perceptions of tests they took so assessors can identify whether the tasks relate to real-world experiences. Alderson et al. (1995) support that test takers would more likely perform better if they consider the tests valid, thus suggesting they would make a stronger effort to perform to the best of their ability when they perceive the tests to accurately measure their language skills. This also suggests a strong correlation between motivation and validity as they relate to level of difficulty.

My questionnaire contained 16 statements with a five-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree), as well as one open-ended question where participants could write their overall opinions about task repetition along with an explanation. This questionnaire was used to determine the ways in which test-takers perceived the tasks (10 statements) and task repetition (6 statements plus 1 open-ended question) in terms of motivation, validity, and difficulty.

Table 3-6 shows the task perception questionnaire statements, most of which were derived from previous questionnaires. The table shows the prior studies that the statements are derived from. This table also indicates what construct was targeted with each statement.

Here I discuss some examples of statements that I derived from previous research, along with reasons why I adapted or adopted them into my questionnaire.

Several examples of Hawkes's (2009) task perception questions that I adapted into the post-task questionnaire for my study were "Do you think that repeating a speaking task is useful for you?", "Did learners feel a sense of achievement after the repeat performance?", and "Do you think it is boring to do the task again?" (Hawkes, 2009, p. 458). I adapted these original questions into the following statements: "Repeating the listening-to-write task was boring" (Statement 2), "Repeating a task helps me improve my writing" (Statement 11), and "Repeating a task helps me improve my listening" (Statement 12). The "useful" and "sense of achievement" aspects in Hawkes's questions could be argued to align with the helpfulness of repeating a task and the improvements that the students made. Therefore, I repurposed these

questions into my own statements to apply to an integrated listening-to-write task repetition. In terms of Hawkes question about it being boring to repeat a task, this aligns with motivation, and some similar relevant questionnaire statements appeared in a task repetition perception questionnaire in Hanzawa and Suzuki's (2023) study: "This task bored me", "I was bored doing this task", and "I would do this task again" (p. 23). These similar task perception questionnaire statements on interest and desire also appeared in Lambert, Philp and Nakamura's (2017) study: "This task bored me", "This task was fun for me", and "I would do this task again" (p. 674).

Some examples of Lofti's (2012) task perception questionnaire statements that I adapted for my questionnaire statements were "I find it difficult to understand well when speakers speak too fast", "I have difficulty understanding a listening text because I cannot understand every single word I hear", and "I find it difficult to do listening tasks, such as filling a grid, for which I need to draw on specific information from the text" (p. 511). I adapted these original statements into the following statements: "The audio recording was played enough times for me to understand it" (Statement 6), "Vocabulary in the audio recording was difficult for me" (Statement 9), and "Writing after listening improves my writing" (Statement 13). In terms of the connection that I have made between Statement 13 and Lofti's statements, I drew from the part of Lofti's statement "…for which I need to draw on specific information from the text". In this case, I was seeking whether the students found that listening helped them with their writing. Because it was after an integrated listening-to-write task repetition that the students completed this questionnaire, it would be possible that repetition helped them gather more input to improve in their writing. Because two of Lofti's statements are about having difficulty understanding a listening text because of not hearing or understanding each word, or possibly pertaining to the speed of speech, I repurposed those statements to align with the possibility that it is necessary for students to hear the audio recording an additional time, thus the reason for Statements 6 and 9. In terms of the statement about vocabulary, Rukthong's (2015) questionnaire also include a similar statement. As for Lofti's statement about the difficulty to draw on listening input in order to complete tasks, I repurposed that statement to align with Statement 13 that also pertains to completing tasks based on listening input.

Next are some examples of task perception questionnaire statements that I adopted from Rukthong's (2015) questionnaire based on a listening-to-write task. Three examples of Rukthong's statements in which the constructs were both validity and difficulty and that I used in my questionnaire were "The task reflects my English writing ability" (Statement 3), The

task reflects my English listening ability" (Statement 4), and "Sentence structures in the audio recording were complicated for me" (Statement 10) (p. 345). Unlike task perception questionnaires from other studies, part of Rukthong's study investigated the listening-to-write task integration, thus I was able to reuse those statements without revising them. These statements also pertained directly to language proficiency in listening and writing in an academic setting where sentence structure was a key component in the study. Therefore, the researcher should know the students' perceptions regarding the tasks so that they can draw meaningful conclusions based on performances as they relate to one another. Two more examples of Rukthong's (2015) statements that I adopted for my questionnaire share a time factor: "I had enough time to complete the writing task" (Statement 5) and "The audio recording was too long" (Statement 8). I included these in my study to capture any patterns in written performances from students who stated that there either was not enough time or that the listening input was too long for them to process.

Although the questionnaire from Rukthong's (2015) study that I discussed above was about a listening-to-write task, I did not use all of the statements that were in Rukthong's questionnaire. Three examples of statements from Rukthong's questionnaire that I did not include in my questionnaire are "Important ideas in the listening passage were paraphrased or repeated more than once", "I could predict the rest of listening content after listening to the first few sentences", and "I had to pay attention to more than one idea at a time" (p. 345). The reasons that I did not include these three are (1) I had already informed the students each time that I would play the listening passage two times, thus I did not need to ask them the number of times I repeated it; (2) because the students heard and read the questions before I played the recording, I anticipated that they would predict the content that would be discussed; and (3) I found that it would be redundant to ask whether they knew that there was more than one idea at a time presented. Further, I was making the length of the questionnaire more manageable, and at the same time, those statements were less directly connected to my research question or constructs of interest.

Two statements that I self-developed for my questionnaire were "I enjoyed repeating the listening-to-write task" (Statement 1) and "Fulfilling a listening-to-write task gets easier with repetition" (Statement 15). In fact, as relevant to the enjoyment statement, in recent research published after I conducted my study, Hanzawa and Suzuki (2023) argued for the need "to obtain further information on learners' emotional engagement with [this task]" (p. 7). Regarding the statement about ease, I wanted to determine after data collection whether

students felt more confident in their language production as a result of the listening-to-write task repetition.

Table 3-6. *Task perception statements derived from previous questionnaires*

		Construct(s) of the Statement			Derived from			Self-Developed
No.	Statement	Motivation	Validity	Difficulty	Rukthong (2015)	Lofti (2012)	Hawkes (2009)	
1	I enjoyed repeating the listening-to-write task	X						X
2	Repeating the listening-to-write task was boring	X					X	
3	The task reflects my English writing ability		X	X	X			
4	The task accurately reflects my English listening ability		X	X	X			
5	I had enough time to complete the writing task		X		X			
6	The audio recording was played enough times for me to understand it		X	X		X		
7	The audio recording provided sufficient ideas for me to complete the writing task		X	X	X			
8	The audio recording was too long	X		X	X	X		
9	Vocabulary in the audio recording was difficult for me	X	X	X	X	X		
10	Sentence structures in the audio recording were complicated for me	X	X	X	X			
11	Repeating a task helps me improve my writing	X	X				X	
12	Repeating a task helps me improve my listening	X	X				X	
13	Writing after listening improves my writing	X		X		X		
14	Listening with the purpose of writing helps me improve my English	X		X	X	X	X	
15	Fulfilling a listening-to-write task gets easier with repetition			X				X

16	I would like to do task repetition in future classes	X		X

In addition to the above 16 Likert-scale statements, the questionnaire included one open-ended question, "What is your overall opinion of task repetition? Explain." Because this question was open-ended, any combination of constructs, i.e., motivation, validity and/or difficulty are applicable. More recent research published after I conducted my study has, in fact, included similar questions. A similar question appeared in Ahmadian et al.'s (2023) task repetition perception questionnaire that the students in their study were provided, i.e., their originally worded question is "How do you feel about repeating a task?" (p. 9). In Ahmadian et al.'s study, there were also two related questions in the task repetition questionnaire that the teachers in their study were asked to answer: "What do you think the purpose of task repetition is?" and "What do you think your students think and feel about task repetition?" (p. 9). In terms of these last two questions, Statements 11 and 12 (see Table 3-6) also align because the goal of improving listening and writing skills in my study are key underpinnings in my task repetition study. The reader would likely infer that these themed task repetition perception questions align somewhat closely with this open-ended question in my study.

3.6 Procedures

Prior to the main study, the instruments and procedures were piloted with the participant population. All of the instruments and procedures were suitable and worked well. Therefore, no changes were needed for the main study

In the main study, the entire data collection process took place in a regular computer room at the university. In the computer room, I alternated the seats so that there was a space between each student. Before the students arrived, I opened the Microsoft Word program, then I deactivated the spelling and grammar autocorrect functions on each computer for this writing activity. Then once the students arrived, and after the information sheets were reviewed with the participants and the signed consent forms were collected, the online biodata survey was administered. After that, all students received their first listening-to-write task. , with no word count minimum or limit. Once the students were finished, they printed their texts from the communal printer and submitted it to me. After the students left the room, I deleted the saved texts so that they could not refer to their previous submission when they returned to repeat it at Time2 and 3. Then, two days later, I provided the feedback group with their performance and my feedback on it.

One week after initial completion, both groups completed their first repetition (Time2), following similar task administration procedures as at Time1, including my repeating the instructions as I did at Time 1. The groups did not have access to their previous written performances at repetition time. The one-week time lapse between Time1 and Time2 was thought to constitute only minimal potential for considerable general language proficiency improvements through other classes or general life experiences. At this point, the students did not know they would get the exact same task again. Then, once the first repetition was completed, the feedback group received feedback two days after the Time2 repetition (as had been the case at Time1).

Finally, two weeks after Time2, both groups repeated the task a final time (Time3). Like at Time2, the groups did not have access to their previous written performances at repetition time. The reason for the two-week interval between Time2 and Time3 was to reduce the possibility that participants would have memorized the input and task, thus potentially skewing the results if improvements were made based on memory rather than skill. At the same time, an even longer interval was not desirable because of potential general language proficiency gains.

The task repetition perception survey was administered after Time3. The reason why I administered it only one time was for practicality; it would have been too time consuming to distribute the questionnaire and collect data after each repetition. Also, it would likely appear tedious for students to receive the same questionnaire, which would likely result in lack of (reliable) responses. Hence, I chose to administer it after the final repetition only, when students had completed all tasks, and may also have formed a more sustained view on task repetition.

3.7 Data Analysis

3.7.1 Analysis of CAF, knowledge summary and transfer scores

To gain initial insights into the nature of the students' listening-to-write performances at each time interval and for each feedback condition, I ran descriptive statistics in SPSS on the scores for each CAF measure and rating scale to determine the mean, standard deviation, minimum, maximum, and range.

Then, to investigate the effects of time and feedback condition (RQ1, RQ2, and RQ3), I ran mixed between-within subjects Analysis of Variance (ANOVA) tests on the scores of each CAF measure and rating scale. This allowed me to determine the statistical significance of any effects of time, feedback condition, and of interactions between time and feedback condition. I

also conducted Mann-Whitney U tests to establish similarity for each measure at Time1 between the two feedback groups. Within my mixed between-within ANOVAs, I conducted a Holm-Bonferroni test for the families of the CAF measures to control for Type 1 errors.

Each ANOVA result answers the following questions:

1. Was there a significant interaction effect between feedback conditions and time periods (repetition) for the types of measures used in the study?
2. Was there a significant main effect of repetition for each measure across the three time points (Time 1-2-3)? Effect sizes were provided. [12]
3. Was there a significant main effect of feedback for each measure (feedback vs. no-feedback) averaged over time points?
4. For measures where there were interaction effects or main effects, what did simple effects tests reveal?

3.7.2 Analysis of task perception questionnaire data

In order to answer RQ4a, students' perceptions of integrated task repetition were investigated by means of responses to the post-task repetition perception questionnaire. Two categories of questions (statements) were posed and examined: a) students' perceptions about the listening-to-write task used in the study and such tasks more generally, and b) students' general perceptions about task repetition. Descriptive statistics were calculated for each statement and for each category to establish the mean, standard deviation, and the minimum and maximum of responses on the five-point Likert scales. Then, to be able to answer RQ4b, first I conducted Kolmogorov-Smirnov tests for normality. The scores for both groups were not normally distributed, and therefore, a non-parametric Mann-Whitney U test was used to assess

---

[1] According to Cohen (1988) and Pallant (2020), an effect size (η_p^2) = .01 is small; an effect size (η_p^2) = .06 is moderate; and an effect size (η_p^2) = .14 or above is large. While there is no strictly defined criterion for the definition of a "very large" effect size, it is worthwhile noting that values much higher than .14 can legitimately occur in repeated-measures designs such as ANOVA (Lakens, 2013). In terms of Cohen's *d,* Cohen's *d* below .20 is a very small effect size; *d* = .20 is a small effect size; *d* = .50 is a medium effect size; and *d* = .80 or above is a large effect size. η_p^2 was used to estimate effect size; Cohen's *d* was used to estimate effect size following simple effects tests.

[2] To provide a more thorough picture of the data, aligning with recommendations for transparency in reporting L2 research data (Plonsky & Oswald, 2014), p-values and effect sizes are reported for both significant and non-significant findings for CAF and knowledge summary and transfer measures.

differences between the groups. Namely, comparative statistics (Mann-Whitney U tests) were run to determine whether the perceptions on the task, on integrated listening-to-write tasks in general, and of task repetition were significantly different between the two feedback conditions.

3.8 Chapter Summary

In this chapter, the overall research design, as well as the data collection setting and techniques were explained. The rationale underlying the data collection techniques was discussed. The overarching question addressed in this study is the impact that integrated listening-to-write task repetition and feedback have on university EAP student writing. Three major research aims were then set to investigate: (1) The extent that task repetition has on CAF and knowledge summary and transfer in writing performances; (2) the extent that feedback affects writing performances; and (3) student perceptions of this task and the degree to which task repetition helps develop their writing proficiency. Sixty-four university EAP students pursuing undergraduate degrees at a university in the United States took part in this study; they were divided into two groups (Feedback and No-Feedback). In order to address the first research aim, the same listening-to-write task texts were collected from the students three different times and analysed quantitatively. To achieve the second research aim, comparative statistics were run to determine whether the feedback group developed better writing performances than did the no-feedback group. To explore the third aim, task perception questionnaire data from the participants were collected and analysed statistically (descriptive and comparative statistics). The next chapter reports the results of the analyses.

4 Results

In this chapter, I first present the results of the analyses conducted to investigate the effects of repetition and feedback condition (RQ1 – interaction effect, RQ2 – repetition effect, and RQ3 – feedback effect) on students' academic writing performances in a listening-to-write task. I describe the results on the CAF measures in section 4.1 and on knowledge summary and transfer (the rating criteria) in section 4.2. For each CAF measure and rating criterion, I first present a table to report the descriptive statistics based on time periods split out by feedback conditions. The population size in all statistics tables where these are split according to feedback condition is N=32 for each group. The raw figures from the descriptive statistics give an initial impression of the direction of any potential repetition and/or feedback effects.

I also show the results from a Mann-Whitney U test comparing the two feedback groups' performances at Time1. Then, I present a table to report the results of the comparative statistical analyses, i.e. mixed-between-within-subjects analysis of variance (ANOVA). Within the comparative statistics tables for each CAF measure, I also include the results from a Holm Bonferroni test to control for Type 1 errors.

In interpreting the findings, attention is also given to whether the changes in performance align with the hypothesised directions for each measure. Each hypothesised direction reflects an expected improvement, expressed as an increase or decrease depending on the measure. For errors per 100 words, errors per T-unit, or in the ratio of simple to complex sentences, decreases indicate improvement; for all other measures, predicted increases indicate improvements. These expected increases or decreases, along with their alignment with the predicted trends, are clarified in the summary tables.

Finally, to gather insights into student perceptions of this task and of task repetition (RQ4a and RQ4b), I ran descriptive statistics on student responses to the perception questionnaire. Then, ANOVAs were used to determine whether perceptions of the task and of task repetition were significantly different between the two feedback conditions. Additionally, I categorized the statements that students made in response to the open-ended question that asked for their overall opinions about repeating a task. The results of these analyses are reported in section 4.3.

4.1 CAF

In this section, I report the results from the analyses of the complexity (4.1.1), accuracy (4.1.2) and (4.1.3), and fluency (4.1.4) of students' written performances on the listening-to-write task.

4.1.1 Complexity

In terms of complexity, the students' written performances on the listening-to-write task were analysed in terms of their: a) sentential complexity, i.e. average sentence length, ratio of simple to complex sentences, and clause per T-unit, b) lexical diversity, i.e. MATTR50, and c) lexical sophistication, i.e. sophisticated words and lexical sophistication proportion. Table 4-1 summarizes the measures used to establish the complexity of the listening-to-write performances, and indicates for each measure what would constitute evidence of increased complexity.

Table 4-1. *Complexity measures*

Construct	Measure	Evidence of improvement	Analysis
Sentential complexity	Average sentence length	Increase in number of words per sentence	MS Word & Manual analysis
	Ratio of simple sentences to complex sentences	Decrease of simple sentences in proportion to complex sentences	Manual analysis
	Average number of clauses per T-unit	Increased number of clauses in proportion to T-units	Manual analysis
Lexical complexity			
Lexical diversity	MATTR50	Higher range of different words within a text	TAALED webtool
Lexical sophistication	Sophisticated words & lexical sophistication	Increased use of sophisticated words & increased number of sophisticated lexical words in proportion to total lexical words	Lextutor lexical analysis webtool

4.1.1.1 Sentential complexity

4.1.1.1.1 Average sentence length

Table 4-2 provides the descriptive statistics for the measure 'average sentence length' in students' writing across time as split out for feedback condition. Students in the feedback condition appeared to have shorter average sentence length at baseline (Time1) compared to students in the no-feedback condition. A Mann-Whitney U test confirmed that there was a statistically significant difference between the two groups' average sentence length at Time1

[U = 349.50, z = -2.185, p = .029], with the no-feedback group having produced on average longer sentences at Time1 than the feedback group. After Time1, the raw figures for the feedback group appear to show a pattern of increase in average sentence length with each repetition. The raw figures for the no-feedback group do not appear to show this pattern. Instead, for students in the no-feedback condition, the raw figures for average sentence length appear to decrease from Time1 to Time2, and then a return to baseline (Time1) at Time3.

Table 4-2. *Descriptive statistics: Average sentence length (by time & feedback condition)*

	Time1		Time2		Time3	
	Feedback Group	No-Feedback Group	Feedback Group	No-Feedback Group	Feedback Group	No-Feedback Group
Mean	15.55	17.71	16.99	14.88	18.18	17.05
SD	7.31	4.42	4.67	2.54	4.54	4.53
Minimum	11.57	8.87	12.44	11.00	11.50	11.68
Maximum	27.60	35.40	24.56	19.64	24.92	25.58

Figure 4.1 visualizes these findings for Time and Feedback group.

Figure 4.1. *Mean: Average sentence length (by feedback condition)*

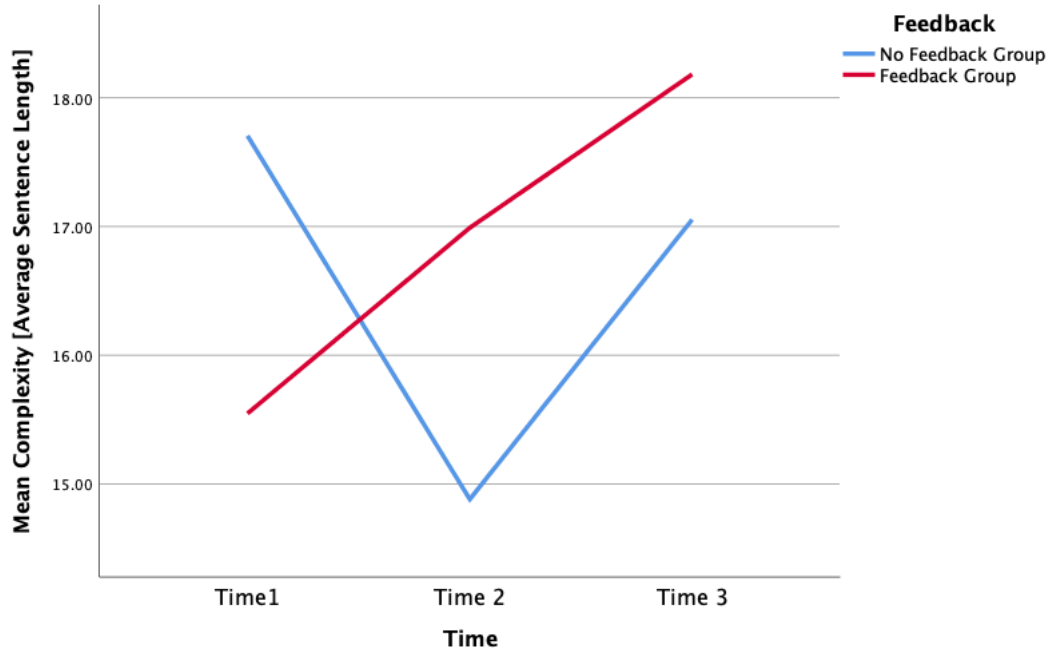


Table 4-3 provides the comparative statistics for average sentence length, including Holm-Bonferroni adjustments to control for Type 1 errors. A 2 (Feedback: Feedback v. No Feedback) x 3 (Time: Time1 v. Time2 v. Time3) between-within subjects Analysis of Variance (ANOVA) was conducted to assess the effects of Feedback and Time (Task Repetition) on average sentence length. This analysis revealed a significant, large interaction effect, [Wilks'

Lambda = 0.82, $F(2, 61) = 6.63$, $p_{adjusted} = .008$, $\eta_p^2 = .18$]. However, this analysis revealed no main effect of Time, [Wilks' Lambda 0.88, $F(2, 61) = 4.11$, $p_{adjusted} = .084$, $\eta_p^2 = .12$], and no main effect of Feedback, [$F(1, 62) = 0.13$, $p_{adjusted} = 2.305$, $\eta_p^2 = .002$].

Table 4-3. *Comparative statistics: Average sentence length*

	ANOVA Test	Unadjusted *p* α=.05	Holm Bonferroni adjusted *p*	Holm Bonferroni adjusted α	Effect Size (η_p^2)	Hypothesised Direction
Time	Wilks' Lambda = 0.88 $F(2, 61) = 4.11$	$p=.021$	$p=.084$	α=.0125		
Feedback condition	$F(1, 62)=.134$	Not significant $p=0.72$	$p=2.305$	α=.025		
Time x Feedback condition	Wilks' Lambda = 0.82 $F(2, 61) = 6.63$	$p<.002$	$p=.008$	α=.0125	Large (.18)	✓

To probe the interaction effect, simple main effects tests were performed, which revealed a significant large main effect of Time for students in the feedback condition, [Wilks' Lambda = 0.67, $F(2, 30) = 7.56$, $p = .002$, $\eta_p^2 = .34$], as well as a significant large main effect of Time for students in the no-feedback condition, [Wilks' Lambda = 0.73, $F(2, 30) = 5.49$, $p = .009$, $\eta_p^2 = .27$].

Between Time1 and Time2, for students who received feedback, simple effects tests revealed that there was no significant difference in average sentence length, $t(31) = -1.95$, $p = .061$, $d = .31$. For students who did not receive feedback, simple effects tests revealed that there was a significant difference in their average sentence length, $t(31) = 2.65$, $p = .01$, $d = .48$, indicating a medium effect, such that average sentence length was significantly shorter at Time2 than Time1.

Between Time2 and Time3, for students who received feedback, simple effects tests revealed that there was no significant difference between average sentence length at Time 2 and Time 3, $t(31) = -1.16$, $p = .255$, $d = .26$. For students who did not receive feedback, simple effects revealed that there was a significant difference, $t(31) = -3.32$, $p = .002$, $d = .59$, indicating a large effect, such that average sentence length at Time3 was significantly longer than Time2.

Between Time1 and Time3, for students who received feedback, simple effects tests revealed that there was a significant difference, $t(31) = -3.52$, $p = .001$, $d = .53$, indicating a large effect, such that average sentence length was significantly longer at Time3 than Time1.

For students who did not receive feedback, simple effects revealed that there was no significant difference, $t(31) = 0.83$, $p = .42$, $d = .10$.

4.1.1.1.2 Ratio of simple sentences to complex sentences

Table 4-4 provides the descriptive statistics for ratio of simple sentences to complex sentences across time as split out for feedback condition. The raw figures suggest that the no-feedback group wrote a lower ratio of simple sentences at Time1 (higher complexity) than the feedback group. A Mann-Whitney U test indicated that there was a statistically significant difference between the two groups' mean ratios of simple/complex sentences at Time 1 [$U = 360.00$, $z = -2.048$, $p = .041$], with the feedback group having produced on average a somewhat higher ratio of simple to complex sentences (lower complexity) in their first performances. The raw data provide an initial indication of the hypothesised pattern in the feedback group, writing increasingly more complex sentences compared to simple sentences (an improvement) with repetitions. The no-feedback group's raw data did not demonstrate this pattern, with similar ratios at baseline (Time1) and Time3, and even a higher ratio of simple sentences (so a potential decline in complexity) at Time2.

Table 4-4. *Descriptive statistics: Ratio: Simple sentence to complex sentence (by feedback condition)*

	Time1		Time2		Time3	
	Feedback Group	No-Feedback Group	Feedback Group	No-Feedback Group	Feedback Group	No-Feedback Group
Mean	1.40	0.96	1.10	1.95	0.91	1.02
SD	0.89	0.74	0.90	3.32	0.51	1.12
Minimum	0.14	0.00	0.00	0.27	0.09	0.09
Maximum	3.00	2.75	3.50	12.00	1.60	4.00

Figure 4.2 visualizes these findings for Time and Feedback group.

Figure 4.2. *Mean: Ratio: Simple to complex sentences (by feedback condition)*

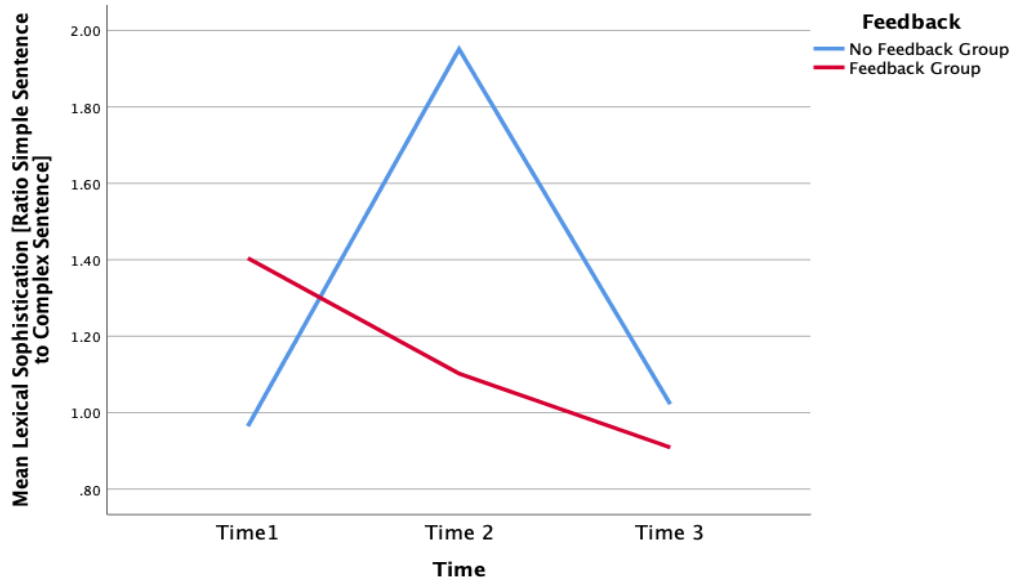


Table 4-5 provides the comparative statistics for the ratio of simple sentences to complex sentences, including Holm-Bonferroni adjustments to control for Type1 errors. A 2 (Feedback: Feedback v. No Feedback) x 3 (Time: Time1 v. Time2 v. Time3) between-within subjects Analysis of Variance (ANOVA) was conducted to assess the effects of Feedback and Time (Task Repetition) on ratio of simple sentences to complex sentences. This analysis revealed a significant, large interaction effect, [Wilks' Lambda = 0.83, $F(2, 61) = 6.33$, $p_{adjusted} = .009$, $\eta_p^2 = .17$]. However, this analysis revealed no main effect of Time, [Wilks' Lambda = 0.94, $F(2, 61) = 1.95$, $p_{adjusted} = .30$, $\eta_p^2 = .06$], and no main effect of Feedback, [$F(1, 62) = 0.4$, $p_{adjusted} = 2.31$, $\eta_p^2 = .01$].

Table 4-5. Comparative statistics: Ratio simple sentence to complex sentence

	ANOVA Test	Unadjusted *p*	Holm Bonferroni adjusted *p*	Holm Bonferroni adjusted α	Effect Size (η_p^2)	Hypothesised Direction
Time	Wilks' Lambda = 0.94 $F(2, 61) = 1.95$	Not significant $p = 0.15$	p=0.30	α=.025		
Feedback condition	$F(1, 62)$=0.4	Not significant $p = 0.53$	p=2.305	α=.0166		
Time x Feedback condition	Wilks' Lambda = 0.83 $F(2, 61) = 6.33$	p =.003	p=.009	α=.0166	Large (.17)	✓

To probe the interaction effect, simple main effects tests were performed, which revealed a significant large main effect of Time for students in the feedback condition, [Wilks' Lambda = 0.67, $F(2,30) = 7.33$, $p = .003$, ηp2 = .33]. However, there was no effect of Time for students in the no-feedback condition, [Wilks' Lambda = 0.87, $F(2,30) = 2.24$, $p = .124$, ηp2 = .13].

Between Time1 and Time2, for students who received feedback, simple effects tests revealed that there was a significant difference in the ratio of simple to complex sentences, $t(31) = 2.30$, $p = .03$, $d = .40$, indicating a small effect, such that the ratio of simple to complex sentences was significantly lower (improved) at Time2 than at Time1. For students who did not receive feedback, simple effects tests revealed that there was no significant difference in the ratio of simple to complex sentences, $t(31) = -2.02$, $p = .052$, $d = .36$.

Between Time2 and Time3, for students who received feedback, simple effects tests revealed that there was also no significant difference in the ratio of simple to complex sentences, $t(31) = 1.20$, $p = .24$, $d = .21$. For students who did not receive feedback, simple effects revealed that there was no significant difference, $t(31) = 1.59$, $p = .12$, $d = .28$.

Between Time1 and Time3, for students who received feedback, simple effects tests revealed that there was a significant difference, $t(31) = -3.66$, $p<.001$, $d = .65$, indicating a medium effect, such that the ratio of simple to complex sentences was significantly lower (an improvement) at Time3 than Time1. For students who did not receive feedback, simple effects revealed that there was no significant difference, $t(31) = -0.26$, $p = .8$, $d = .45$.

4.1.1.1.3 Clauses per T-unit [C/T]

Table 4-6 provides the descriptive statistics for the average number of C/T in students' writing across time as split out for feedback condition. The no-feedback group's performances started at a higher average number of clauses per T-unit than the Feedback group (Time1). A Mann-Whitney U test indicated that there was a statistically significant difference between the two groups' average C/T at Time1. [$U = 337.50$, $z = -2.352$, $p = .019$], with the no-feedback group having produced on average more C/T in their first performances. The raw figures suggest rather stable average C/T numbers within each group across repetitions; the means remain pretty similar in each group over time.

Table 4-6. *Descriptive statistics: C/T (by feedback condition)*

	Time1		Time2		Time3	
	Feedback Group	No-Feedback Group	Feedback Group	No-Feedback Group	Feedback Group	No-Feedback Group
Mean	1.31	1.50	1.29	1.43	1.33	1.45

Table 4-6. *Descriptive statistics: C/T (by feedback condition)*

	Time1		Time2		Time3	
	Feedback Group	No-Feedback Group	Feedback Group	No-Feedback Group	Feedback Group	No-Feedback Group
SD	0.19	0.29	0.43	0.37	0.13	0.24
Minimum	1.00	1.22	0.59	0.85	1.14	1.11
Maximum	1.60	2.00	3.13	2.17	1.50	1.85

Figure 4.3 visualizes these findings for Time and Feedback group.

Figure 4.3. *Mean: C/T (by feedback condition)*

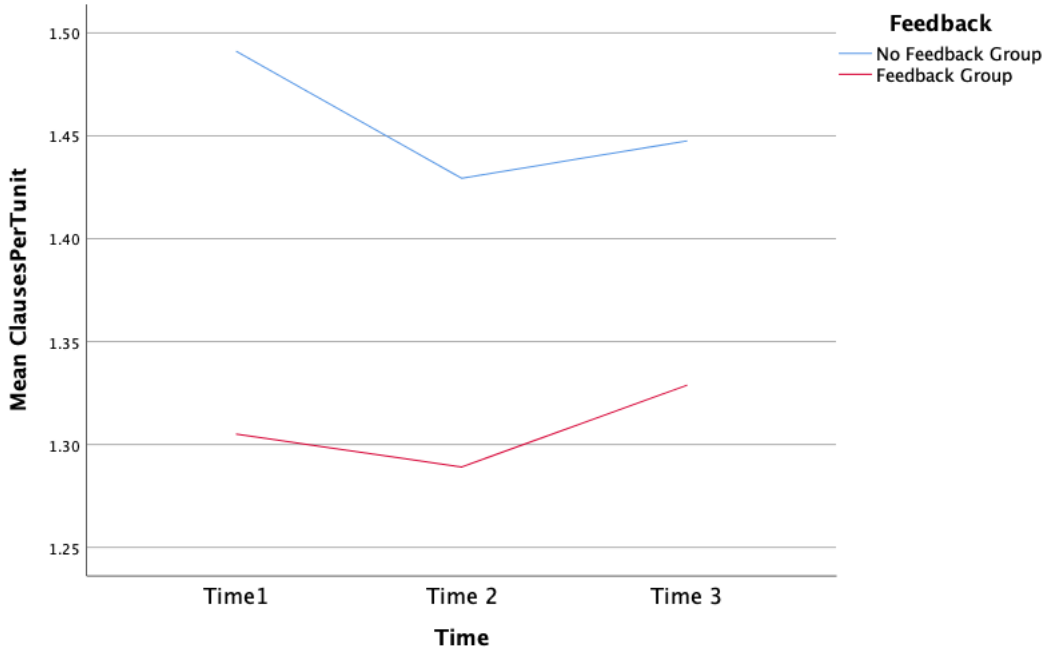


Table 4-7 provides the comparative statistics for average number of clauses per T-unit, including Holm-Bonferroni adjustments to control for Type 1 errors. A 2 (Feedback: Feedback v. No Feedback) x 3 (Time: Time1 v. Time2 v. Time3) between-within subjects Analysis of Variance (ANOVA) was conducted to assess the effects of Feedback and Time (Task Repetition) on C/T. This analysis revealed no significant interaction effect, [Wilks' Lambda = 0.98, $F(2,61)=.71$, $p_{adjusted}=.495$, $\eta_p^2=.02$], as well as no significant main effect of Time [Wilks' Lambda = 0.99, $F(2,61)=.22$, $p_{adjusted}=.81$, $\eta_p^2=.01$]. Also, there was no significant main effect of Feedback, $F(1,62)=10.22$, $p_{adjusted=}.012$, $\eta_p^2=.14$].

Table 4-7. *Comparative Statistics: C/T*

	ANOVA Test	Unadjusted *p*	Holm Bonferroni adjusted *p*	Holm Bonferroni adjusted α	Effect Size (η_p^2)	Hypothesised Direction
Time	Wilks' Lambda = 0.99 $F(2, 61) = 0.22$	Not significant $p=.081$	$p=0.81$	α=.0166		

Feedback condition	$F(1, 62)=10.22$	$p=.002$	$p=.012$	α=.0083
Time x Feedback condition	Wilks' Lambda = 0.98 $F(2, 61) = .711$	Not significant $p=.495$	$p=.495$	α=0.05

4.1.1.2 Lexical complexity

4.1.1.2.1 Lexical diversity

Table 4-8 provides the descriptive statistics for lexical diversity in students' writing across time as split out for feedback condition. The raw figures suggest that the no-feedback group had slightly higher lexical diversity than the no-feedback group at Time1. A Mann-Whitney U test indicated that there was no statistically significant difference between the two groups' average lexical diversity at Time1. [$U = 414.00$, $z = - 1.322$, $p = .186$]. The feedback group's raw lexical diversity data increased slightly at Time2 and Time3, whereas they remained the same for the no-feedback group at Time2 and just slightly increased at Time3.

Table 4-8. *Descriptive statistics: Lexical diversity (by feedback condition)*

	Time1		Time2		Time3	
	Feedback Group	No-Feedback Group	Feedback Group	No-Feedback Group	Feedback Group	No-Feedback Group
Mean	0.677	0.700	0.697	0.699	0.716	0.714
SD	0.061	0.040	0.064	0.039	0.056	0.039
Minimum	0.580	0.630	0.600	0.640	0.600	0.660
Maximum	0.760	0.790	0.780	0.790	0.810	0.790

Figure 4.4 visualizes these findings for Time and Feedback group.

Figure 4.4. *Mean: Lexical diversity (by feedback condition)*

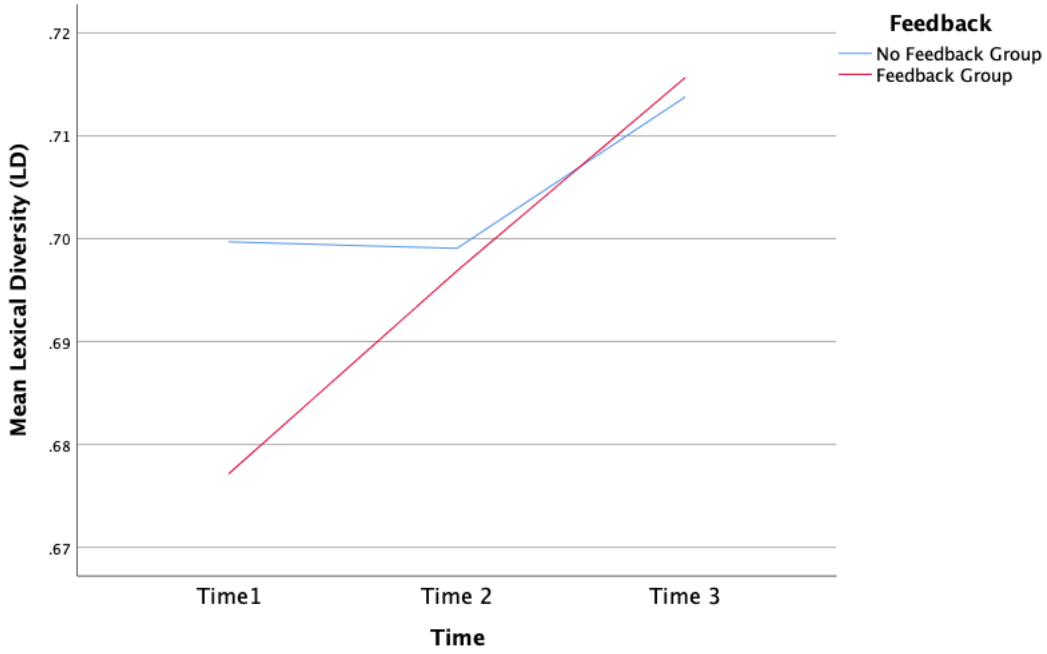


Table 4-9 provides the comparative statistics for lexical diversity, including Holm-Bonferroni adjustments to control for Type 1 errors. A 2 (Feedback: Feedback v. No Feedback) x 3 (Time: Time1 v. Time2 v. Time3) between-within subjects Analysis of Variance (ANOVA) was conducted to assess the effects of Feedback and Time (Task Repetition) on lexical diversity. This analysis revealed no significant interaction effect, [Wilks' Lambda = 0.92, $F(2, 61) = 2.65$, $p_{adjusted}$=..158, η_p^2=.08], as well as no significant main effect of Feedback, [$F(1, 62) = 0.48$, $p_{adjusted}$=2.305, η_p^2=.01]. However, this analysis revealed a large main effect of Time, [Wilks' Lambda = 0.77, $F(2, 61) = 9.23$, $p_{adjusted}$<.001, η_p^2 = .23].

Table 4-9. *Comparative statistics: Lexical diversity*

	ANOVA Test	Unadjusted *p*	Holm Bonferroni adjusted *p*	Holm Bonferroni adjusted α	Effect Size (η_p^2)	Hypothesised Direction
Time	Wilks' Lambda = 0.77 $F(2, 61) = 9.23$	p<.001	p<.001	α=.0083	Large (.23)	✓
Feedback condition	$F(1, 62)$= 0.48	Not significant p=.493	p=2.305	α=.0125		
Time x Feedback condition	Wilks' Lambda = 0.92 $F(2, 61) = 2.65$	Not significant p=.079	p=.158	α=.025		

Simple effects tests revealed that there was no significant difference in lexical diversity between Time1 and Time2, $t(63) = -1.87$, $p = .066$, $d = .23$, However, there was a significant difference between Time2 and Time3, $t(63) = -3.07$, $p=.003$, $d = .38$, as well as a significant difference between Time1 and Time3, $t(63)= -4.21$, $p<.001$, $d = .53$.

4.1.1.2.2 Lexical sophistication [number of sophisticated words]

Table 4-10 provides the descriptive statistics for the mean number of sophisticated words in students' writing across time as split out for feedback condition. The raw figures suggest that the feedback group started at a lower average at Time1 than the no-feedback group. A Mann-Whitney U test indicated that there was a statistically significant difference between the two groups' mean numbers of sophisticated words at Time1. [$U = 313.00$, $z = -2.702$, $p = .007$], with the no-feedback group having produced on average more sophisticated words in their first performances. The raw figures suggest that both groups follow the pattern of increases at each time. The no-feedback group made only a slight increase from Time2 to Time3 while the feedback group made steadier increases at each time.

Table 4-10. *Descriptive statistics: Sophisticated words (by feedback condition)*

	Time1		Time2		Time3	
	Feedback Group	No-Feedback Group	Feedback Group	No-Feedback Group	Feedback Group	No-Feedback Group
Mean	5.97	7.44	7.13	10.13	14.34	10.16
SD	2.15	2.85	2.61	5.14	6.53	4.71
Minimum	3.00	2.00	4.00	4.00	8.00	3.00
Maximum	10.00	11.00	13.00	22.00	26.00	16.00

Figure 4.5 visualizes these findings for Time and Feedback group.

Figure 4.5. *Mean: Sophisticated words*

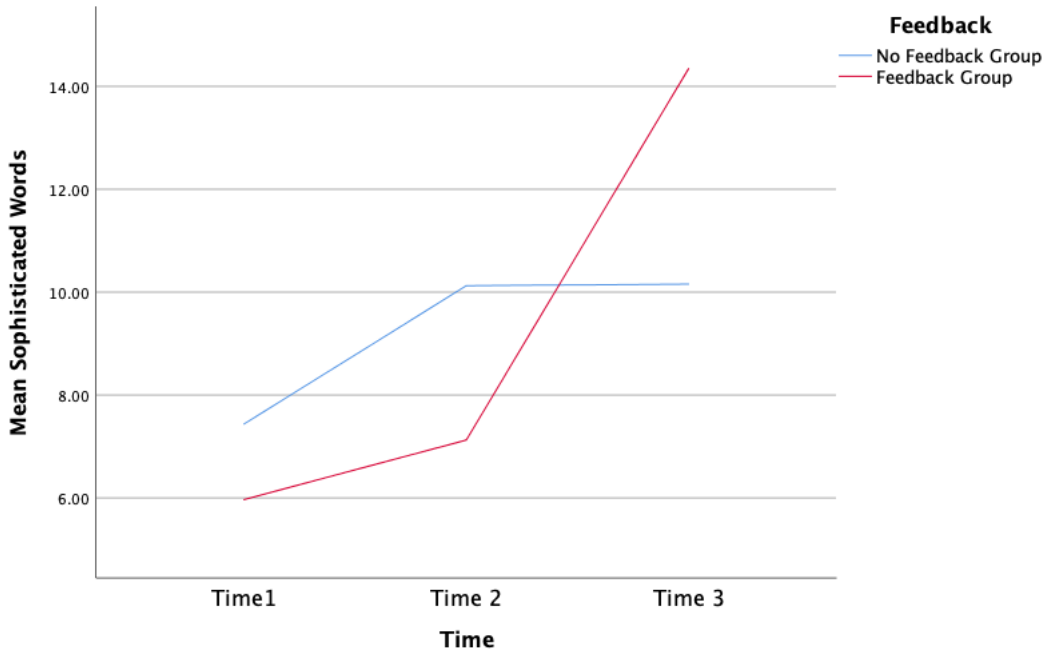


Table 4-11 provides the comparative statistics for the mean number of sophisticated words, including Holm-Bonferroni adjustments to control for Type1 errors. A 2 (Feedback: Feedback v. No Feedback) x 3 (Time: Time1 v. Time2 v. Time3) between-within subjects Analysis of Variance (ANOVA) was conducted to assess the effects of Feedback and Time (Task Repetition) on sophisticated words. This analysis revealed a significant, large interaction effect, [Wilks' Lambda = 0.72, $F(2, 61) = 12.18$, $p_{adjusted}<.001$ $\eta_p^2=.29$]. This analysis also revealed a large main effect of Time, [Wilks' Lambda = 0.42, $F(2, 61) = 42.56$, $p_{adjusted}<.001$ $\eta_p^2=.58$], i.e., the students wrote more sophisticated words on average after repetitions However, the analysis revealed there was no main effect of Feedback, [$F(1, 62) = 0.14$, $p_{adjusted} = 2.31$, $\eta_p^2=.000$].

Table 4-11. *Comparative statistics: Sophisticated words*

	ANOVA Test	Unadjusted *p*	Holm Bonferroni adjusted *p*	Holm Bonferroni adjusted α	Effect Size (η_p^2)	Hypothesised Direction
Time	Wilks' Lambda = 0.42 $F(2, 61) = 42.56$	$p<.001$	$p<.001$	α=0.01	Large (.58)	✓
Feedback condition	$F(1, 62)= .014$	Not significant $p = .91$	$p=2.31$	α=0.05		
Time x Feedback condition	Wilks' Lambda = 0.72 $F(2, 61) = 12.18$	$p<.001$	$p<.001$	α=.0083	Large (.29)	✓

To probe the interaction effect, simple main effects tests were performed, which revealed a significant large main effect of Time for students in the feedback condition, [Wilks' Lambda = 0.35, $F(2,30) = 28$, $p<.001$, ηp2 = .65]. There was also a significant large main effect of Time for students in the no-feedback condition, [Wilks' Lambda $F(2,30) = 0.48$, $p<.001$, ηp2 = .52].

Between Time1 and Time2, for students who received feedback, simple effects tests revealed that there was a significant difference in the average number of sophisticated words, $t(31) = -4.54$, $p<.001$, $d = .80$, indicating a large effect, such that the average number of sophisticated words was significantly higher at Time2 than at Time1. For students who did not receive feedback, simple effects tests revealed that there was a significant difference Time1 and Time2, $t(31) = -3.03$, $p = .005$, $d = .54$, indicating a medium effect, such that the average number of sophisticated words was significantly higher at Time2 than at Time1.

Between Time2 and Time3, for students who received feedback, simple effects tests revealed that there was also a significant difference in the average number of sophisticated words, $t(31) = -6.55$, $p<.001$, $d = 1.16$, indicating a large effect, such that the average number of sophisticated words was significantly higher at Time3 than at Time2. For students who did not receive feedback, simple effects revealed that there was no significant difference, $t(31) = -0.03$, $p = .98$, $d = .01$.

Between Time1 and Time3, for students who received feedback, simple effects tests revealed that there was a significant difference, $t(31) = -7.24$, $p<.001$, $d = 1.28$, indicating a large effect, such that the average number of sophisticated words was significantly higher at Time3 than at Time1. For students who did not receive feedback, simple effects revealed that there was a significant difference, $t(31) = -5.25$, $p<.001$, $d = .93$, such that the average number of sophisticated words was significantly higher at Time3 than at Time1.

4.1.1.2.3 Lexical sophistication proportion [sophisticated lexical words in proportion to total lexical words]

Table 4-12 provides the descriptive statistics for lexical sophistication proportion [sophisticated lexical words in proportion to total lexical words] in students' writing across time as split out for feedback condition. The raw figures suggest that students in the Feedback and No-Feedback conditions started with similar lexical sophistication means at baseline (Time1). A Mann-Whitney U test confirmed that there was no significant difference between the two groups' average lexical sophistication at Time1 [$U = 419.00$, $z = -1.273$, $p = .203$], with both groups having scored on average quite similarly on lexical sophistication in their

first performances. The raw figures suggest that the feedback group had quite similar means at Time1 and Time 2, then an increase at Time3; the no-feedback group had an increase from Time1 to Time2, then a decrease at Time3.

Table 4-12. *Descriptive statistics: Lexical sophistication [proportion] (by feedback condition)*

	Time1		Time2		Time3	
	Feedback Group	No-Feedback Group	Feedback Group	No-Feedback Group	Feedback Group	No-Feedback Group
Mean	0.045	0.051	0.041	0.059	0.057	0.043
SD	0.02	0.02	0.01	0.03	0.02	0.02
Minimum	0.02	0.01	0.02	0.02	0.03	0.01
Maximum	0.07	0.08	0.08	0.11	0.11	0.08

Figure 4.6 visualizes these findings for Time and Feedback group.

Figure 4.6. *Mean: Lexical sophistication [proportion] (by feedback condition)*

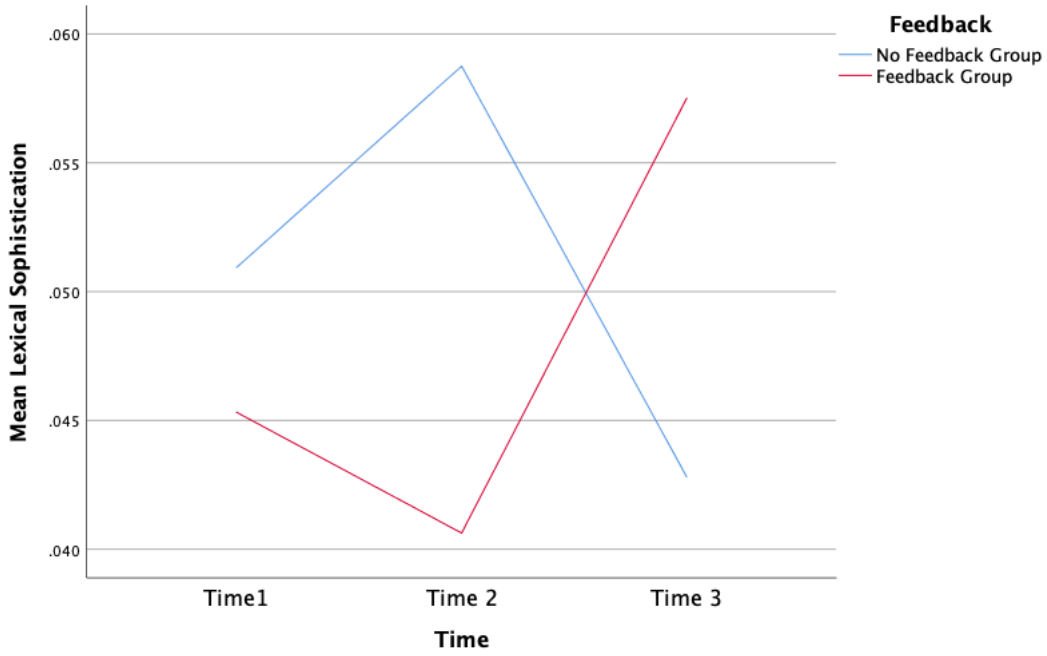


Table 4-13 provides the comparative statistics for lexical sophistication, including Holm-Bonferroni adjustments to control for Type 1 errors. A 2 (Feedback: Feedback v. No Feedback) x 3 (Time: Time1 v. Time2 v. Time3) between-within subjects Analysis of Variance (ANOVA) was conducted to assess the effects of Feedback and Time (Task Repetition) on lexical sophistication. This analysis revealed a significant, large interaction effect, [Wilks' Lambda = 0.63, $F(2, 61) = 17.77$, $p_{adjusted}<.001$, $\eta_p^2=.37$]. However, this analysis revealed no main effect of Time, [Wilks' Lambda = 0.99, $F(2, 61) = 0.35$, $p_{adjusted} = .705$, $\eta_p^2=.01$], and no main effect of Feedback, [$F(1, 62) = 0.55$, $p_{adjusted} = 2.305$, $\eta_p^2=.01$].

Table 4-13. *Comparative statistics: Lexical sophistication [proportion]*

	ANOVA Test	Unadjusted *p*	Holm Bonferroni adjusted *p*	Holm Bonferroni adjusted α	Effect Size (η_p^2)	Hypothesised Direction
Time	Wilks' Lambda = 0.99 *F*(2, 61) = 0.35	Not significant *p*=.705	*p*=.705	α=0.05		
Feedback condition	*F*(1, 62)=0.55	Not significant *p*=.461	*p*=2.305	α=0.01		
Time x Feedback condition	Wilks' Lambda = 0.63 *F*(2, 61) = 17.77	*p*<.001	*p*<.001	α=0.01	Large (.37)	✓

To probe the interaction effect, simple main effects tests were performed, which revealed a significant large main effect of Time for students in the feedback condition, [Wilks' Lambda = 0.75, $F(2,30) = 9.93$, $p<.001$, ηp2 = .25], as well as a significant large main effect of Time for students in the no-feedback condition, [Wilks' Lambda = 0.79, $F(2,30) = 8.2$, $p<.001$, ηp2 = .21].

Between Time1 and Time2, for students who received feedback, simple effects tests revealed that there was no significant difference in lexical sophistication, $t(31) = 1.49$, $p = .15$, $d = .26$. Similarly, for students who did not receive feedback, simple effects tests revealed that there was no significant difference, $t(31) = -1.91$, $p = .07$, $d = .34$,

Between Time2 and Time3, for students who received feedback, simple effects tests revealed that there was a significant difference in lexical sophistication, $t(31) = -3.68$, $p<.001$, $d = .65$, indicating a medium effect, such that average lexical sophistication at Time3 was significantly higher than Time2. For students who did not receive feedback, simple effects revealed that there was a significant difference, $t(31) = 5.19$, $p<.001$, $d = .92$, indicating a large effect, such that average lexical sophistication at Time3 was significantly lower than Time2.

Between Time1 and Time3, for students who received feedback, simple effects tests revealed that there was a significant difference, $t(31) = -2.93$, $p = .006$, $d = .52$, indicating a medium effect, such that average lexical sophistication was significantly higher at Time3 than Time1. For students who did not receive feedback, simple effects revealed that there was a significant difference, $t(31) = 2.68$, $p = .01$, $d = .48$, indicating a small effect, such that average lexical sophistication was significantly lower at Time3 than Time1.

4.1.2 Accuracy (by global measure)

In this section, I show the results for the accuracy of the students' written linguistic performances on the listening-to-write task, analysed in terms of errors per 100 words, errors per T-unit (E/T) and error-free T-units per T-unit (EFT/T). Accuracy is looked at here for the combination of all types of grammatical errors taken together, as specified in Table 4-14. The table also summarizes the measures used to establish the accuracy of the listening-to-write performances, and clarifies what constitutes an improvement in accuracy. Then in the section after, I show the results split out for individual types of grammatical errors (e.g. verb tense errors).

Table 4-14. *Accuracy measures*

Construct	Measure	Evidence of improvement	Analysis
Grammar: • Subject-verb agreement • Verb tense • Verb form usage • Prepositions • Articles	Errors per 100 words	Decreased number of errors per 100 words	Manual analysis
	Errors per T-unit (E/T)	Decreased average number of errors per T-unit	Manual analysis
	Error-free T-units per T-unit (EFT/T)	Increased average number of error-free T-units per T-unit	Manual analysis

4.1.2.1 Errors per 100 words

Table 4-15 provides the descriptive statistics for the average number of errors per 100 words in students' writing across time as split out for feedback condition. The feedback group's performances contained a somewhat higher average number of errors per 100 words at Time1 than the no-feedback group. However, a Mann-Whitney U test indicated that there was no statistically significant difference between the two groups' mean errors per 100 words at Time1. [$U = 379.50$, $z = -1.782$, $p = .075$], with both groups having scored on average quite similarly on errors per 100 words in their first performances. The raw figures suggest both groups follow a pattern of making fewer errors per 100 words at each time (so, an improvement in accuracy). The feedback group made a steadier decrease from Time1 to Time2 than the no-feedback group. Then, from Time2 to Time3, both groups made somewhat similar decreases, averaging similar means at Time3.

Table 4-15. *Descriptive statistics: Errors per 100 words (by feedback condition)*

	Time1		Time2		Time3	
	Feedback Group	No-Feedback Group	Feedback Group	No-Feedback Group	Feedback Group	No-Feedback Group
Mean	5.73	4.78	4.42	4.28	2.55	2.75
SD	2.59	1.84	1.63	1.87	1.54	0.97
Minimum	2.94	1.61	1.69	2.31	1.30	0.69
Maximum	12.35	8.16	7.14	7.44	6.58	3.77

Figure 4.7 visualizes these findings for Time and Feedback group.

Figure 4.7. *Mean: Errors per 100 words (by feedback condition)*

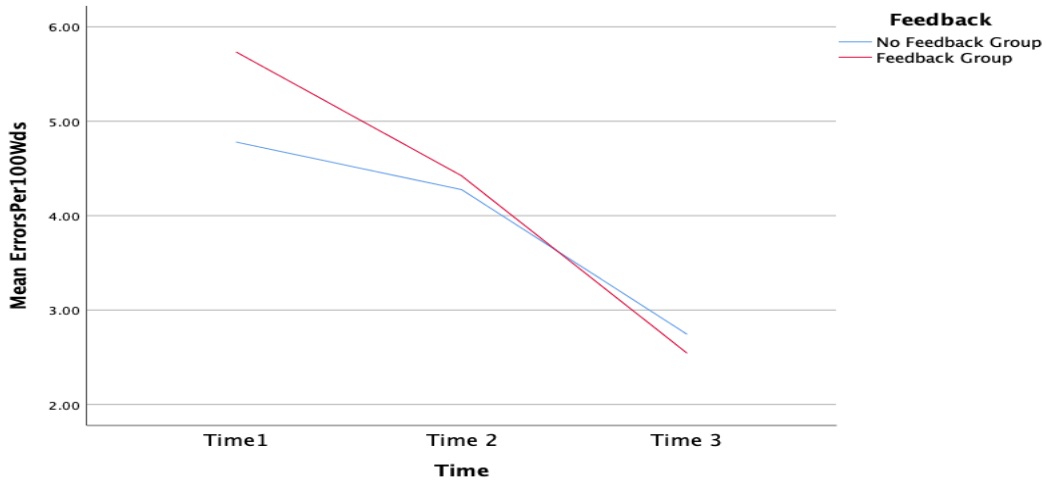


Table 4-16 provides the comparative statistics for average number of errors per 100 words, including Holm-Bonferroni adjustments to control for Type 1 errors. A 2 (Feedback: Feedback v. No Feedback) x 3 (Time: Time 1 v. Time 2 v. Time 3) between-within subjects Analysis of Variance (ANOVA) was conducted to assess the effects of Feedback and Time (Task Repetition) on students' average number of errors per 100 words. This analysis revealed no significant interaction effect, [Wilks' Lambda = 0.91, $F(2, 61) = 3.18$, *padjusted* = .147, η_p^2=.09], as well as no significant main effect of Feedback, [$F(1, 62) = 0.75$, $p = 1.17$, η_p^2 = .01]. However, this analysis revealed a large main effect of Time, [Wilks' Lambda = 0.25, $F(2, 61) = 93.93$, p<.001, η_p^2 = .76].

Table 4-16. *Comparative statistics: Errors per 100 words*

	ANOVA Test	Unadjusted *p*	Holm Bonferroni adjusted *p*	Holm Bonferroni adjusted α	Effect Size (η_p^2)	Hypothesised Direction
Time	Wilks' Lambda = 0.25 $F(2, 61) = 93.93$	p<.001	p<.001	α=.0166	Large (.76)	✓
Feedback condition	$F(1, 62)$= 0.75	Not significant p=0.39	p=1.17	α=.0166		

Time x Feedback condition	Wilks' Lambda = 0.91 $F(2, 61) = 3.18$	p=.049	p=.147	α=.0166

Simple effects tests revealed a significant difference in average number of errors per 100 words between Time 1 and Time 2, $t(63) = 2.94$, $p = .005$, $d = .37$, a significant difference between Time 2 and Time 3, $t(63) = 8.8$, $p<.001$, $d = 1.1$, as well as a significant difference between Time 1 and Time 3, $t(63)= 10.45$, $p<.001$, $d = 1.31$.

4.1.2.2 Errors per T-Unit (E/T)

Table 4-17 provides the descriptive statistics for E/T in students' writing across time as split out for feedback condition. The raw figures suggest that the feedback group had a somewhat lower average number of E/T than the no-feedback group at Time1. However, a Mann-Whitney U test indicated that there was no statistically significant difference between the two groups' mean E/T at Time 1. [$U = 496.50$, $z = -.209$, $p = .835$], with both groups having scored on average quite similarly on E/T in their first performances. The raw figures suggest that both groups follow a pattern of making fewer E/T at each time (so, an improvement in accuracy). The no-feedback group made a steadier decrease (an improvement) from Time1 to Time2 and from Time2 to Time3.

Table 4-17. *Descriptive statistics: E/T (by feedback condition)*

	Time1		Time2		Time3	
	Feedback Group	No-feedback Group	Feedback Group	No-feedback Group	Feedback Group	No-feedback Group
Mean	0.73	0.82	0.60	0.57	0.38	0.37
SD	0.33	0.63	0.27	0.25	0.19	0.13
Minimum	0.29	0.20	0.24	0.31	0.13	0.10
Maximum	1.43	2.60	1.00	1.08	0.71	0.54

Figure 4.8 visualizes these findings for Time and Feedback group.

Figure 4.8. *Mean: E/T (by feedback condition)*

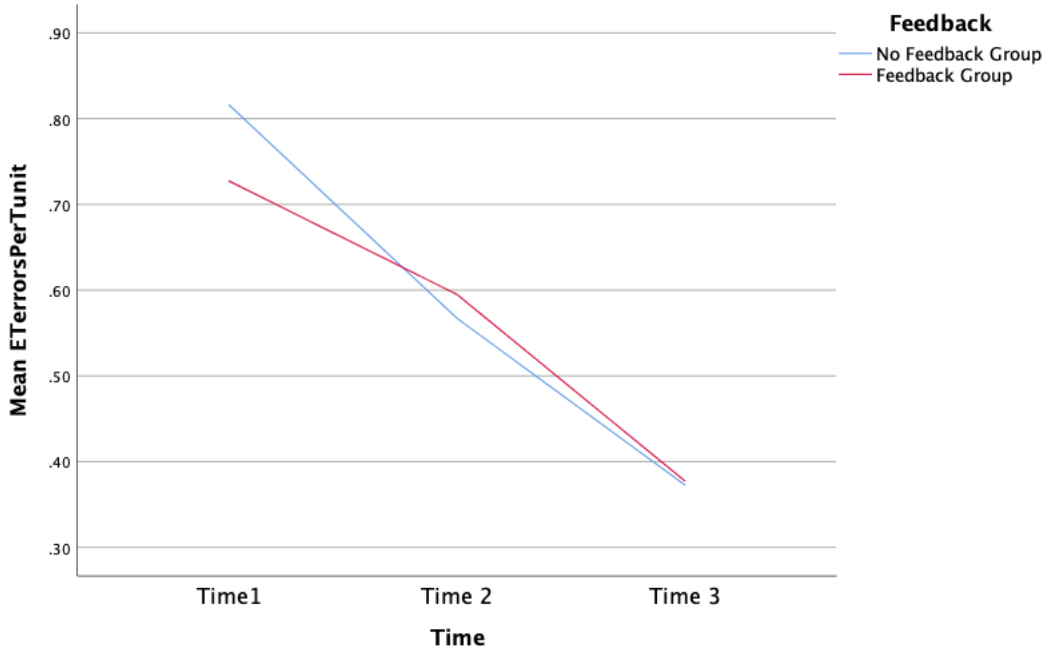


Table 4-18 provides the comparative statistics for E/T, including Holm-Bonferroni adjustments to control for Type 1 errors. A 2 (Feedback: Feedback v. No Feedback) x 3 (Time: Time1 v. Time2 v. Time3) between-within subjects Analysis of Variance (ANOVA) was conducted to assess the effects of Feedback and Time (Task Repetition) on E/T. This analysis revealed no significant interaction effect, [Wilks' Lambda = 0.98, $F(2, 61) = 0.51$, $p_{adjusted} = .794$, $\eta_p^2 = .02$], as well as no significant main effect of Feedback, [$F(1, 62) = 0.10$, $p_{adjusted} = 1.17$, $\eta_p^2 = .002$]. However, this analysis revealed a large main effect of Time, [Wilks' Lambda = 0.40, $F(2, 61) = 44.99$, $p_{adjusted} < .001$, $\eta_p^2 = .60$].

Table 4-18. *Comparative statistics: E/T*

	ANOVA Test	Unadjusted *p*	Holm Bonferroni adjusted *p*	Holm Bonferroni adjusted α	Effect Size (η_p^2)	Hypothesised Direction
Time	Wilks' Lambda = 0.40 $F(2, 61) = 44.99$	$p<.001$	$p<.001$	α=.025	Large (.60)	✓
Feedback condition	$F(1, 62)= 0.10$	Not significant $p=0.75$	$p=1.17$	α=0.05		
Time x Feedback condition	Wilks' Lambda = 0.98 $F(2, 61) = 0.51$	Not significant $p=0.60$	$p=.794$	α=0.05		

Simple effects tests revealed that there was a significant difference in E/T between Time1 and Time2, $t(63) = 3.094$, $p = .003$, $d = .39$, a significant difference between Time2 and

Time3, $t(63) = 8.701$, $p<.001$, $d = .1088$, as well as a significant difference between Time1 and Time3, $t(63)= 6.263$, $p<.001$, $d = .783$.

4.1.2.3 Error-Free T-Units per T-Unit

Table 4-19 provides the descriptive statistics for EFT/T in students' writing across time as split out for feedback condition. The raw figures suggest that both groups had similar means of EFT/T at Time1 and Time2. A Mann-Whitney U test indicated that there was no statistically significant difference between the two groups' mean EFT/T at Time1. [$U = 434.50$, $z = -1.055$, $p = .291$], with both groups having scored on average quite similarly on EFT/T in their first performances. Then, both groups improved with very similar numbers of EFT/T at Time3.

Table 4-19. *Descriptive statistics: EFT/T (by feedback condition)*

	Time1		Time2		Time3	
	Feedback Group	No-feedback Group	Feedback Group	No-feedback Group	Feedback Group	No-feedback Group
Mean	0.66	0.60	0.66	0.67	0.76	0.75
SD	0.15	0.20	0.15	0.24	0.16	0.09
Minimum	0.29	0.22	0.44	0.15	0.54	0.62
Maximum	0.86	0.87	0.88	0.90	1.00	0.95

Figure 4.9 visualizes these findings for Time and Feedback group.

Figure 4.9. *Mean: EFT/T (by feedback condition)*

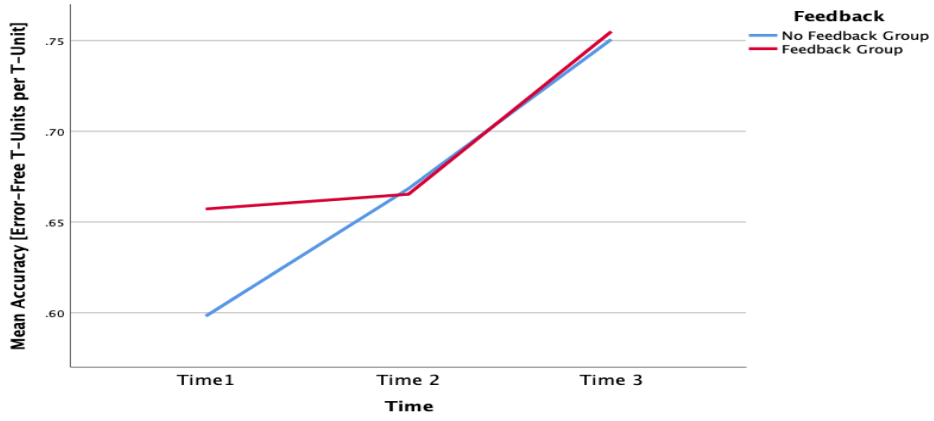


Table 4-20 provides the comparative statistics for EFT/T, including Holm-Bonferroni adjustments to control for Type 1 errors. A 2 (Feedback: Feedback v. No Feedback) x 3 (Time: Time1 v. Time2 v. Time3) between-within subjects Analysis of Variance (ANOVA) was conducted to assess the effects of Feedback and Time (Task Repetition) on EFT/T. This analysis revealed no significant interaction effect, [Wilks' Lambda = 0.97, $F(2, 61) = .94$, $p_{adjusted} = .794$, $\eta_p^2=.03$], as well as no significant main effect of Feedback, [$F(1, 62) = 0.38$,

$p_{adjusted}$ = 1.17, η_p^2 = .006]. However, this analysis revealed a large main effect of Time, [Wilks' Lambda = 0.66, $F(2, 61) = 15.43$, $p_{adjusted}$<.001, η_p^2 = .34].

Table 4-20. *Comparative statistics: EFT/T*

	ANOVA Test	Unadjusted *p*	Holm Bonferroni adjusted *p*	Holm Bonferroni adjusted α	Effect Size (η_p^2)	Hypothesised Direction
Time	Wilks' Lambda = 0.66 $F(2, 61) = 15.43$	p<.001	p<.001	α=0.05	Large (.34)	✓
Feedback condition	$F(1, 62)$=0.38	Not significant p=.539	p=1.17	α=.025		
Time x Feedback condition	Wilks' Lambda = 0.97 $F(2, 61) = 0.94$	Not significant p=.397	p=.794	α=.025		

Simple effects tests revealed that there was no significant difference in E/T between Time1 and Time2, $t(63) = -1.57$, $p = .123$, $d = .20$. However, there was a significant difference between Time2 and Time3, $t(63) = -3.74$, p<.001, $d = .47$, as well as a significant difference between Time1 and Time3, $t(63)= -5.36$, p<.001, $d = .67$.

4.1.3 Accuracy (by error type)

In this section, I show the results for the accuracy measure 'error type per 100 words' by error type, i.e., subject-verb agreement, verb tense, verb form, prepositions, and articles.

4.1.3.1 Subject-verb agreement errors per 100 words

Table 4-21 provides the descriptive statistics for the average number of subject-verb agreement errors per 100 words in students' writing across time as split out for feedback condition. The raw figures suggest that the no-feedback group started at higher average numbers of subject-verb agreement errors per 100 words than the feedback group at Time1. A Mann-Whitney U test indicated that there was no statistically significant difference between the two groups' mean number of sophisticated words at Time1 [$U = 506.50$, $z = -.075$, $p = .941$], with both groups having scored on average quite similarly on subject-verb agreement errors per 100 words in their first performances. The raw figures for both groups follow the pattern of decreases (so, an improvement in accuracy) at each time. The feedback group made only a very slight decrease (very slight improvement) from Time1 to Time2, and a steadier decrease (an improvement) from Time2 to Time3. The no-feedback group made steady decreases (improvements) at each time.

Table 4-21. *Descriptive statistics: Subject verb agreement errors per 100 words (by feedback condition)*

	SVA Errors per 100 Words Time1		SVA Errors per 100 Words Time2		SVA Errors per 100 Words Time3	
	Feedback Group	No-Feedback Group	Feedback Group	No-Feedback Group	Feedback Group	No-Feedback Group
Mean	1.333	1.632	1.316	1.040	0.482	0.459
SD	1.135	1.760	0.997	1.206	0.427	0.416
Minimum	0.000	0.000	0.000	0.000	0.000	0.000
Maximum	3.700	6.120	3.770	3.550	1.350	1.250

Figure 4.10 visualizes these findings for Time and Feedback group.

Figure 4.10. *Mean: Subject verb agreement errors per 100 words (by feedback condition)*

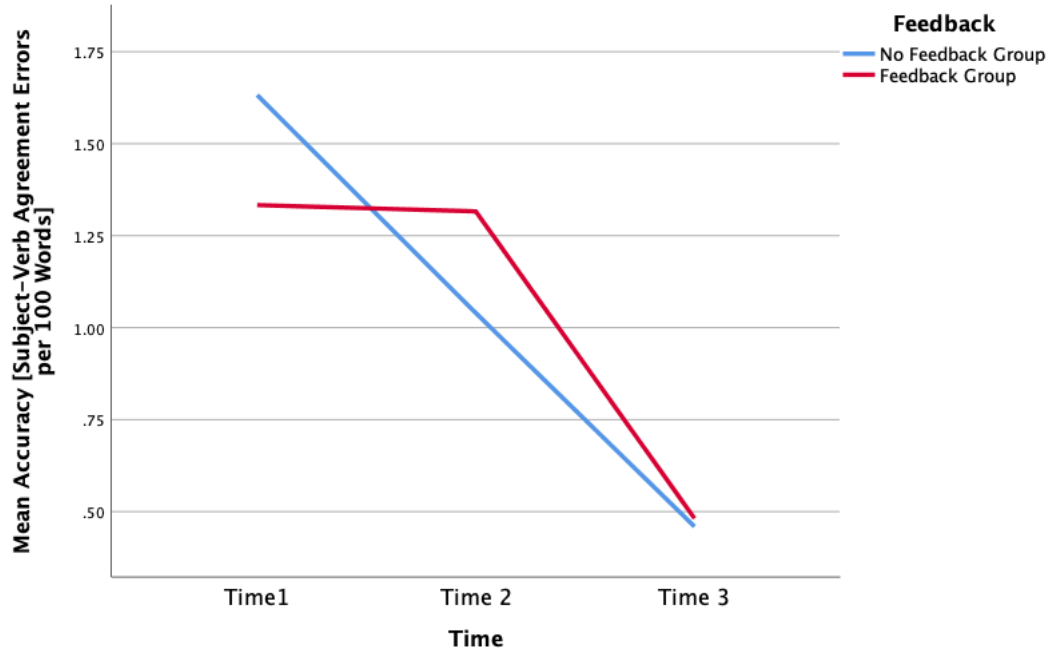


Table 4-22 provides the comparative statistics for average number of subject-verb agreement errors per 100 words, including Holm-Bonferroni adjustments to control for Type 1 errors. A 2 (Feedback: Feedback v. No Feedback) x 3 (Time: Time1 v. Time2 v. Time3) between-within subjects Analysis of Variance (ANOVA) was conducted to assess the effects of Feedback and Time (Task Repetition) on subject-verb agreement errors. This analysis revealed no significant interaction effect, [Wilks' Lambda = 0.97, $F(2, 61) = 1.07$, *padjusted* = .698, η_p^2=.03], as well as no significant main effect of Feedback, [$F(1, 62) = .000$, $p_{adjusted}$ = .999, $\eta_p^2 = .000$]. However, this analysis revealed a large main effect of Time, [Wilks' Lambda = 0.48, $F(2, 61) = 33.57$, $p_{adjusted}$<.001, $\eta_p^2 = .52$].

Table 4-22. *Comparative statistics: Subject-Verb Agreement Errors per 100 Words*

	ANOVA Test	Unadjusted *p*	Holm Bonferroni adjusted *p*	Holm Bonferroni adjusted α	Effect Size (η_p^2)	Hypothesised Direction
Time	Wilks' Lambda = 0.48 *F*(2, 61) = 33.57	*p*<.001	*p*<.001	α=.0125	Large (.52)	✓
Feedback condition	*F*(1, 62)=.000	Not significant *p*=.999	*p*=.999	α=0.05		
Time x Feedback condition	Wilks' Lambda = 0.97 *F*(2, 61) = 1.07	Not significant *p*=.349	*p*=.698	α=.025		

Simple effects tests revealed that there was no significant difference in average subject-verb agreement errors per 100 words between Time1 and Time2, $t(63) = 1.522$, $p = .133$, $d = .19$. However, there was a significant difference between Time2 and Time3, $t(63) = 5.83$, $p<.001$, $d = .73$, as well as a significant difference between Time1 and Time3, $t(63)= 6.167$, $p<.001$, $d = .77$.

4.1.3.2 Verb tense errors per 100 words

Table 4-23 provides the descriptive statistics for average number of verb tense errors per 100 words across time as split out for feedback condition. The raw data suggest that the feedback group started at a higher average number of verb tense errors per 100 words than the no-feedback group in their first performances based on the raw data, but a Mann-Whitney U test indicated that there was no statistically significant difference between the two groups' mean number of verb tense errors per 100 words at Time1. [$U = 609.00$, $z = 1.466$, $p = .143$], with both groups having scored on average quite similarly on verb tense errors per 100 words in their first performances. The raw figures suggest that the feedback group had a minimal increase in errors from Time1 to Time2, then a steady decrease (i.e., an improvement) from Time2 to Time3. The no-feedback group made a slight decrease (a slight improvement) from Time1 to Time2, then an increase from Time2 to Time3 where the average number of errors was slightly higher than at baseline (Time1).

Table 4-23. *Descriptive statistics: Verb tense errors per 100 words (by feedback condition)*

	Verb Tense Errors per 100 Words Time1		Verb Tense Errors per 100 Words Time2		Verb Tense Errors per 100 Words Time3	
	Feedback Group	No-Feedback Group	Feedback Group	No-Feedback Group	Feedback Group	No-Feedback Group
Valid	32	32	32	32	32	32
Mean	0.420	0.253	0.436	0.222	0.234	0.256

Table 4-23. *Descriptive statistics: Verb tense errors per 100 words (by feedback condition)*

	Verb Tense Errors per 100 Words Time1		Verb Tense Errors per 100 Words Time2		Verb Tense Errors per 100 Words Time3	
	Feedback Group	No-Feedback Group	Feedback Group	No-Feedback Group	Feedback Group	No-Feedback Group
SD	0.481	0.375	0.425	0.323	0.245	0.365
Minimum	0.000	0.000	0.000	0.000	0.000	0.000
Maximum	1.230	1.020	1.070	0.830	0.660	1.030

Figure 4.11 visualizes these findings for Time and Feedback group.

Figure 4.11. *Mean: Verb tense errors per 100 words*

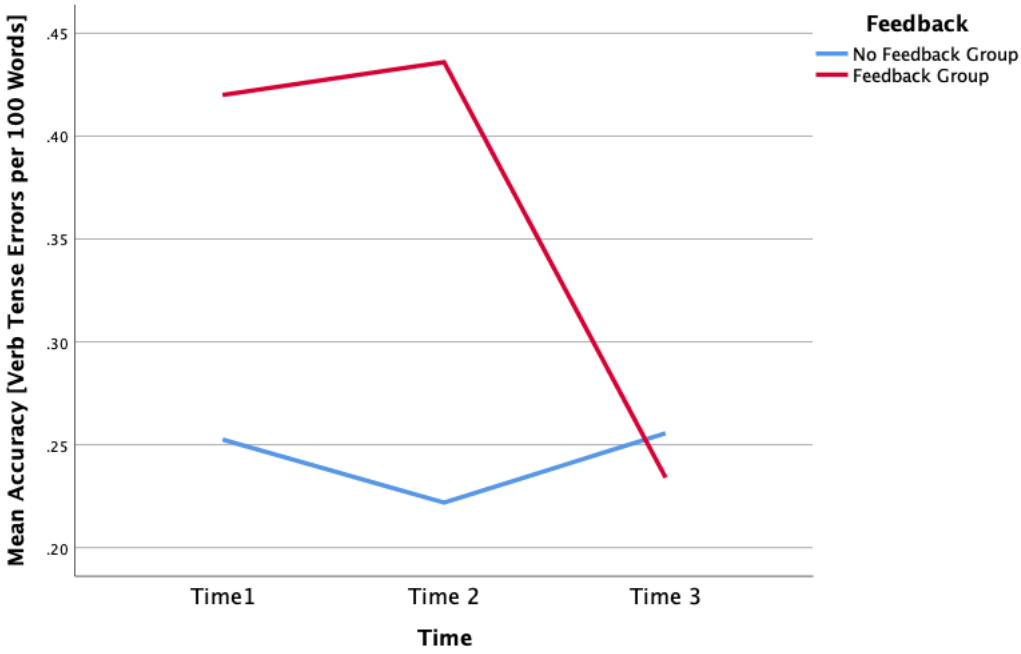


Table 4-24 provides the comparative statistics for average number of verb tense errors per 100 words, including Holm-Bonferroni adjustments to control for Type 1 errors. A 2 (Feedback: Feedback v. No Feedback) x 3 (Time: Time1 v. Time2 v. Time3) between-within subjects Analysis of Variance (ANOVA) was conducted to assess the effects of Feedback and Time (Task Repetition) on verb tense errors per 100 words. This analysis revealed no significant interaction effect, [Wilks' Lambda = 0.93, $F(2,61) = 2.26$, $p_{adjusted} = .452$, $\eta_p^2 = .07$], as well as no significant main effect of Time [Wilks' Lambda = 0.96, $F(2,61) = 1.34$, $p_{adjusted} = .27$, $\eta_p^2 = .04$]. Also, there was no significant main effect of Feedback, [$F(1,62) = 4.3$, $p_{adjusted} = .168$, $\eta_p^2 = .07$].

Table 4-24. *Comparative Statistics: Verb tense per 100 words*

	ANOVA Test	Unadjusted *p*	Holm Bonferroni adjusted *p*	Holm Bonferroni adjusted α	Effect Size (η_p^2)	Hypothesised Direction

Time	Wilks' Lambda = 0.96 $F(2, 61) = 1.34$	Not significant $p=.27$	$p=0.27$	α=0.05
Feedback condition	$F(1, 62)=4.31$	$p=.042$	$p=.168$	α=.0125
Time x Feedback condition	Wilks' Lambda = 0.93 $F(2, 61) = 2.26$	Not significant $p=.113$	$p=.452$	α=.0125

4.1.3.3 Verb form errors per 100 words

Table 4-25 provides the descriptive statistics for the average number of verb form errors per 100 words in students' writing across time as split out for feedback condition. The raw figures suggest that the feedback group started with a higher average number of verb form errors per 100 words than the no-feedback group at Time1, although a Mann-Whitney U test indicated that there was no statistically significant difference between the two groups' mean number verb form errors per 100 words at Time1. [$U = 507.50$, $z = -.061$, $p = .952$]. The raw figures for both groups follow the pattern of decreases at each time (so, accuracy improvements). The feedback group made steady decreases (improvements) at each time. The no-feedback group made a slight decrease (improvement) from Time1 to Time2, then a steadier decrease (improvement) from Time2 to Time3.

Table 4-25. *Descriptive statistics: Mean verb form errors per 100 words (feedback condition)*

	Verb Form Errors per 100 Words Time1		Verb Form Errors per 100 Words Time2		Verb Form Errors per 100 Words Time3	
	Feedback Group	No-Feedback Group	Feedback Group	No-Feedback Group	Feedback Group	No-Feedback Group
Mean	1.884	1.694	1.536	1.614	1.134	1.251
SD	1.421	1.003	0.891	0.926	1.142	0.648
Minimum	0.760	0.000	0.000	0.580	0.000	0.340
Maximum	4.940	3.000	3.570	3.310	3.950	2.090

Figure 4.12 visualizes these findings for Time and Feedback group.

Figure 4.12. *Mean: Verb form errors per 100 words (by feedback condition)*

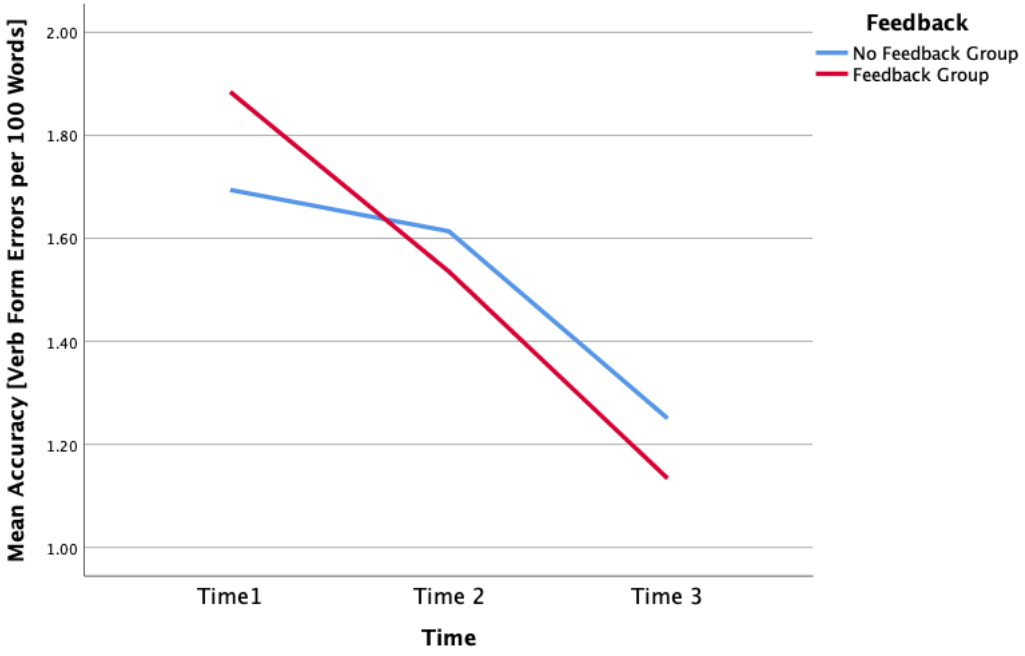


Table 4-26 provides the comparative statistics for average number of verb form errors per 100 words, including Holm-Bonferroni adjustments to control for Type 1 errors. A 2 (Feedback: Feedback v. No Feedback) x 3 (Time: Time1 v. Time2 v. Time3) between-within subjects Analysis of Variance (ANOVA) was conducted to assess the effects of Feedback and Time (Task Repetition) on verb form errors. This analysis revealed no significant interaction effect, [Wilks' Lambda = 0.98, $F(2, 61) = 0.54$, $p_{adjusted} = .698$, $\eta_p^2 = .02$], as well as no significant main effect of Feedback, [$F(1, 62) = .000$, $p_{adjusted} = 1.988$, $\eta_p^2 = .000$]. However, this analysis revealed a large main effect of Time, [Wilks' Lambda = 0.78, $F(2, 61) = 8.69$, $p_{adjusted} < .001$, $\eta_p^2 = .22$], i.e., the students made fewer verb form errors per 100 words (so, accuracy improvements) on average after repetitions.

Table 4-26. *Comparative Statistics: Verb Form Errors per 100 Words*

	ANOVA Test	Unadjusted *p*	Holm Bonferroni adjusted *p*	Holm Bonferroni adjusted α	Effect Size (η_p^2)	Hypothesised Direction
Time	Wilks' Lambda = 0.78 $F(2, 61) = 8.69$	$p<.001$	$p<.001$	α=0.01	Large (.22)	✓
Feedback condition	$F(1, 62)=.000$	Not significant $p=.994$	$p=1.988$	α=.025		
Time x Feedback condition	Wilks' Lambda = 0.98 $F(2, 61) = 0.54$	Not significant $p=.584$	$p=.698$	α=0.05		

Simple effects tests revealed that there was no significant difference in average verb form errors per 100 words between Time1 and Time2, $t(63) = .636$, $p = .107$, $d = .20$. However, there was a significant difference between Time2 and Time3, $t(63) = 3.83$, $p<.001$, $d = .48$, as well as a significant difference between Time1 and Time3, $t(63)= 3.69$, $p<.001$, $d = .46$.

4.1.3.4 Preposition errors per 100 words

Table 4-27 provides the descriptive statistics for the average number of preposition errors per 100 words in students' writing across time as split out for feedback condition. The raw figures suggest that students in the feedback group started at a higher average number of preposition errors per 100 words at Time1 than the no-feedback group. A Mann-Whitney U test indicated that there was a statistically significant difference between the two groups' mean number of preposition errors per 100 words at Time1, [$U = 773.50$, $z = 3.538$, $p>.001$], with the no-feedback group having produced on average fewer preposition errors per 100 words than the feedback group in their first performances. The raw data suggest that the feedback group made steady decreases (so, accuracy improvements) at each time. The no-feedback group made an increase (so, a decline) at Time2, then a steady decrease (so, an improvement) at Time3.

Table 4-27. *Descriptive statistics: Number of preposition errors per 100 words (by feedback condition)*

	Preposition errors per 100 words Time 1		Preposition errors per 100 words Time 2		Preposition errors per 100 words Time 3	
	Feedback Group	No-Feedback Group	Feedback Group	No-Feedback Group	Feedback Group	No-Feedback Group
Mean	1.629	0.770	0.924	0.900	0.442	0.470
SD	0.942	0.712	0.790	0.440	0.333	0.332
Minimum	0.000	0.000	0.000	0.410	0.000	0.000
Maximum	2.940	2.480	2.420	2.110	1.080	0.920

Figure 4.13 visualizes these findings for Time and Feedback group.

Figure 4.13. *Mean: Preposition errors per 100 words*

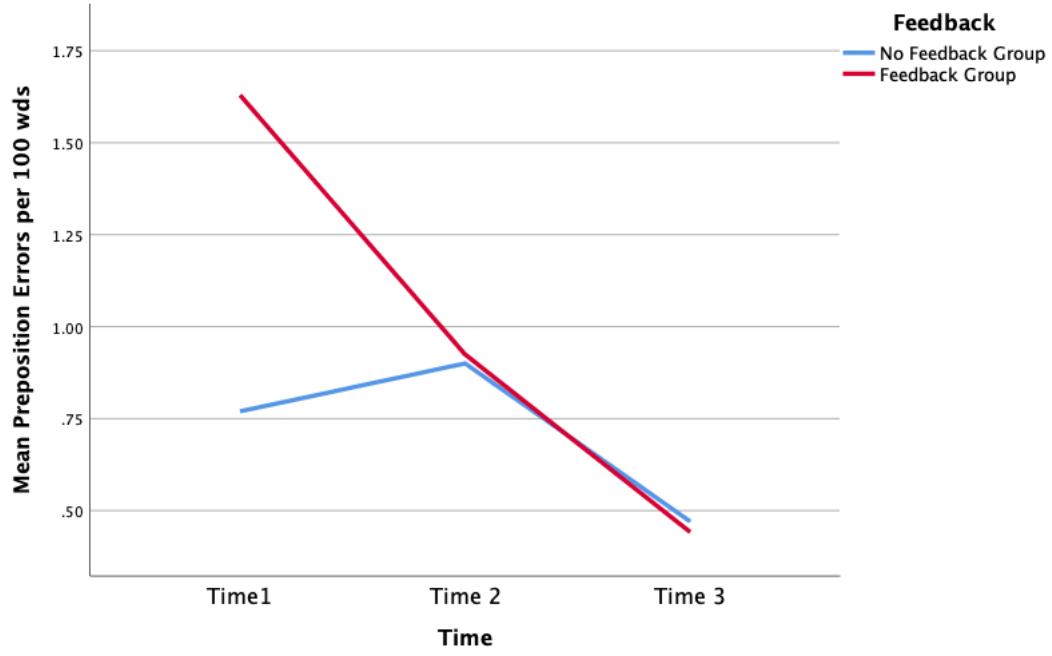


Table 4-28 provides the comparative statistics for average number of preposition errors per 100 words, including Holm-Bonferroni adjustments to control for Type 1 errors. A 2 (Feedback: Feedback v. No Feedback) x 3 (Time: Time1 v. Time2 v. Time3) between-within subjects Analysis of Variance (ANOVA) was conducted to assess the effects of Feedback and Time (Task Repetition) on preposition errors. This analysis revealed a significant, large interaction effect, [Wilks' Lambda = 0.75, $F(2, 61) = 10.21$, $p_{adjusted}<.001$, $\eta_p^2=.25$]. This analysis revealed a large main effect of Time, [Wilks' Lambda = 0.45, $F(2, 61) = 37.9$, $p_{adjusted}<.001$, $\eta_p^2=.55$], , i.e. the students made fewer preposition errors per 100 words on average after repetitions. However, there was no main effect of Feedback, [$F(1, 62) = 5.66$, $p_{adjusted} = .10$, $\eta_p^2 = .08$].

Table 4-28. *Comparative Statistics: Preposition errors per 100 words*

	ANOVA Test	Unadjusted *p*	Holm Bonferroni adjusted *p*	Holm Bonferroni adjusted α	Effect Size (η_p^2)	Hypothesised Direction
Time	Wilks' Lambda = 0.45 $F(2, 61) = 37.9$	$p<.001$	$p<.001$	α=.0167	Large (.55)	✓
Feedback condition	$F(1, 62)= 5.66$	$p=0.02$	$p=0.10$	α=0.01		
Time x Feedback condition	Wilks' Lambda = 0.75 $F(2, 61) = 10.21$	$p<.001$	$p<.001$	α=0.01	Large (.25)	✓

To probe the interaction effect, simple main effects tests were performed, which revealed a significant large main effect of Time for students in the feedback condition, [Wilks' Lambda = 0.34, $F(2,30) = 28.74$, $p<.001$, $\eta p2 = .66$], as well as a significant large main effect of Time for students in the no-feedback condition, [Wilks' Lambda = 0.50, $F(2,30) = 15.02$, $p<.001$, $\eta p2 = .50$].

Between Time1 and Time2, for students who received feedback, simple effects tests revealed that there was a significant difference in preposition errors per 100 words, $t(31) = 4.17$, $p<.001$, $d = .74$, indicating a medium effect, such that average number of preposition errors per 100 words at Time 2 was significantly lower (an accuracy improvement) than Time1. For students who did not receive feedback, simple effects tests revealed that there was no significant difference, $t(31) = -1.19$, $p = .25$, $d = .21$.

Between Time2 and Time3, for students who received feedback, simple effects tests revealed that there was a significant difference between average number of preposition errors per 100 words, $t(31) = 4.46$, $p<.001$, $d = .79$, indicating a medium effect, such that average number of preposition errors at Time 3 was significantly lower (an improvement) than Time2. For students who did not receive feedback, simple effects revealed that there was a significant difference, $t(31) = 5.57$, $p<.001$, $d = .98$, indicating a large effect, such that average number of preposition errors at Time3 was significantly lower (an improvement) than Time2.

Between Time1 and Time3, for students who received feedback, simple effects tests revealed that there was a significant difference, $t(31) = 7.27$, $p<.001$, $d = 1.29$, indicating a large effect, such that average number of preposition errors was significantly lower (an improvement) at Time3 than Time1. For students who did not receive feedback, simple effects revealed that there was a significant difference, $t(31) = 2.56$, $p = .02$, $d = .45$, indicating a small effect, such that average number of preposition errors was significantly lower (an improvement) at Time3 than Time1.

4.1.3.5 Article errors per 100 words

Table 4-29 provides the descriptive statistics for average number of article errors per 100 words in students' writing across time as split out for feedback condition. The feedback group started at a slightly higher average number of article errors per 100 words than the no-feedback group. However, a Mann-Whitney U test indicated that there was no statistically significant difference between the two groups' mean number of sophisticated words at Time1 [$U = 505.00$, $z = -.099$, $p = .921$], with both groups having scored on average quite similarly on article errors per 100 words in their first performances. The raw numbers suggest that the

feedback group made a steady decrease from Time1 to Time2 (an improvement), then a slight increase at Time3 (a decline). The raw numbers suggest that the no-feedback group made an increase at Time2 (a decline), then a steady decrease at each Time3 (an improvement).

Table 4-29. *Descriptive statistics: Number of article errors per 100 words (by feedback condition)*

	Article errors per 100 words Time1		Article errors per 100 words Time2		Article errors per 100 words Time3	
	Feedback Group	No-Feedback Group	Feedback Group	No-Feedback Group	Feedback Group	No-Feedback Group
Mean	0.464	0.430	0.210	0.498	0.258	0.305
SD	0.625	0.417	0.305	0.721	0.451	0.303
Minimum	0.000	0.000	0.000	0.000	0.000	0.000
Maximum	2.040	1.420	0.850	2.340	1.450	0.680

Figure 4.14 visualizes these findings for Time and Feedback group.

Figure 4.14. Mean: Article errors per 100 words (by feedback condition)

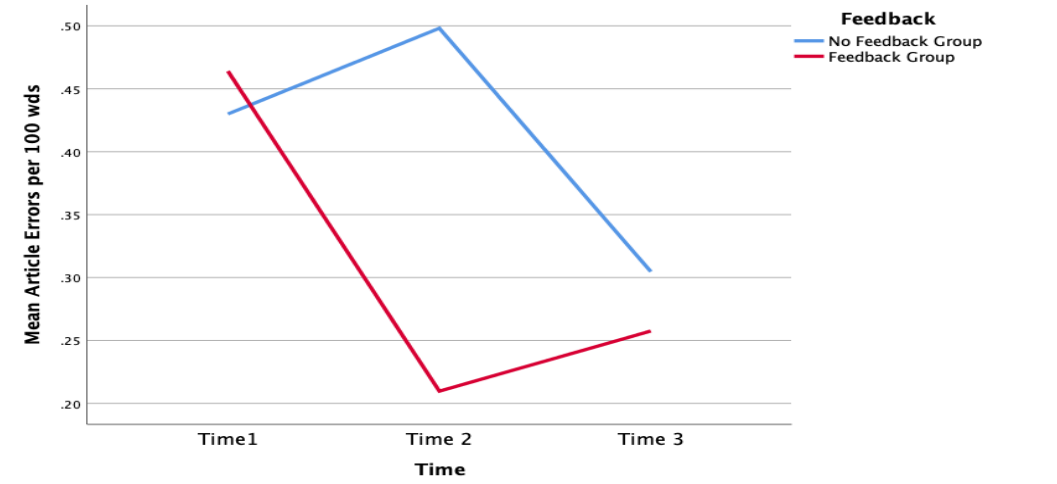


Table 4-30 provides the comparative statistics for average number of article errors per 100 words, including Holm-Bonferroni adjustments to control for Type 1 errors. A 2 (Feedback: Feedback v. No Feedback) x 3 (Time: Time1 v. Time2 v. Time3) between-within subjects Analysis of Variance (ANOVA) was conducted to assess the effects of Feedback and Time (Task Repetition) on students' average number of article errors per 100 words. This analysis revealed no significant interaction effect, [Wilks' Lambda = 0.94, $F(2,61)=1.99$, $p_{adjusted}=.452$, $\eta_p^2=.06$], as well as no significant main effect of Time [Wilks' Lambda = 0.90, $F(2,61)=3.29$, $p_{adjusted}=.088$, $\eta_p^2=.10$]. Also, there was no significant main effect of Feedback, [$F(1,62)=1.29$, $p_{adjusted}=.78$, $\eta_p^2 = .02$].

Table 4-30. *Comparative statistics: Article errors per 100 Words*

	ANOVA Test	Unadjusted *p*	Holm Bonferroni adjusted *p*	Holm Bonferroni adjusted α	Effect Size (η_p^2)	Hypothesised Direction
Time	Wilks' Lambda = 0.90 *F*(2, 61) = 3.29	*p*=.044	*p*=.088	α=.025		
Feedback condition	*F*(1, 62)=1.29	Not significant *p*=.260	*p*=0.78	α=.0167		
Time x Feedback condition	Wilks' Lambda = 0.94 *F*(2, 61) = 1.99	Not significant *p*=.145	*p*=.452	α=.0167		

4.1.4 Fluency

The fluency measures of the students' written performances on the listening-to-write task were: Words per T-unit [W/T] and words per error-free T-unit [W/EFT]. Table 4-31 defines the measures and indicates what constitutes an improvement in written fluency.

Table 4-31. *Fluency measures*

Construct	Measure	Evidence of improvement	Analysis
Written language proficiency	Words per T-unit (W/T)	Increased average number of words per T-unit	Manual analysis
	Words per error-free T-unit (W/EFT)	Increased average number of words per error-free T-unit	Manual analysis

4.1.4.1 Words per T-Unit (W/T)

Table 4-32 provides the descriptive statistics for W/T across time as split out for feedback condition. The raw figures suggest that students in the no-feedback group started at a higher mean number of W/T at Time1 than the feedback group, but a Mann-Whitney U test indicated that there was no statistically significant difference between the two groups' mean number of W/T at Time1. [$U = 408.00$, $z = -1.398$, $p = .162$], with both groups having scored on average quite similarly on W/T in their first performances. The figures suggest that the feedback group follows the expected pattern of more W/T with repetition. The no-feedback group deviates from this by starting with higher W/T, and then levelling off with fewer W/T at Time2 and Time3.

Table 4-32. *Descriptive statistics: W/T (by feedback condition)*

	Time1		Time2		Time3	
	Feedback Group	No-Feedback Group	Feedback Group	No-Feedback Group	Feedback Group	No-Feedback Group
Mean	12.52	16.07	13.24	13.51	15.16	13.84
SD	1.93	7.19	2.51	2.51	4.86	2.64
Minimum	9.50	10.23	9.65	9.73	9.00	10.07
Maximum	16.73	35.40	17.67	16.57	24.92	18.46

Figure 4.15 visualizes these findings for Time and Feedback group.

Figure 4.15. *Mean: W/T (by feedback condition)*

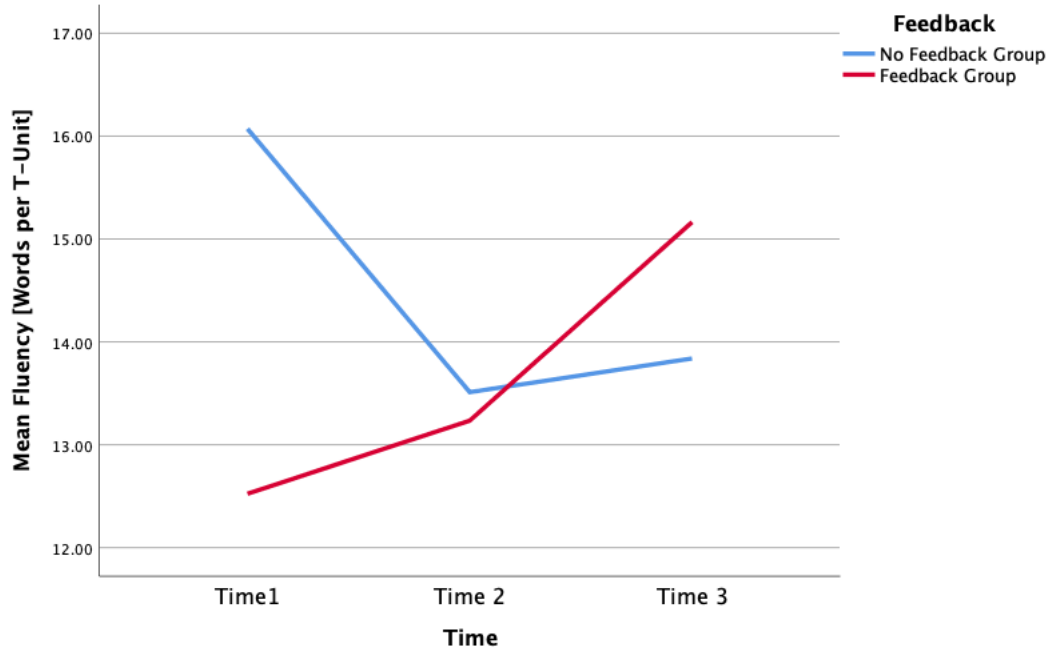


Table 4-33 provides the comparative statistics for mean number of W/T, including Holm-Bonferroni adjustments to control for Type 1 errors. A 2 (Feedback: Feedback v. No Feedback) x 3 (Time: Time1 v. Time2 v. Time3) between-within subjects Analysis of Variance (ANOVA) was conducted to assess the effects of Feedback and Time (Task Repetition) on W/T. This analysis revealed a significant, large interaction effect, [Wilks' Lambda = 0.80, $F(2, 61) = 7.66$, $p_{adjusted}<.001$, $\eta_p^2=.20$]. This analysis revealed no main effect of Time, [Wilks' Lambda = 0.92, $F(2, 61) = 2.64$, $p_{adjusted} = .08$, $\eta_p^2=.08$]. Also, there was no main effect of Feedback, [$F(1, 62) = 1.21$, $p_{adjusted} = .275$, $\eta_p^2 = .02$].

Table 4-33. *Comparative statistics: W/T*

	ANOVA Test	Unadjusted *p*	Holm Bonferroni adjusted *p*	Holm Bonferroni adjusted α	Effect Size (η_p^2)	Hypothesised Direction

Time	Wilks' Lambda = 0.92 $F(2, 61) = 2.64$	Not significant p=0.08	p=0.08	α=0.05		
Feedback condition	$F(1, 62)$=1.21	Not significant p=.275	p=.275	α=0.05		
Time x Feedback condition	Wilks' Lambda =0.80 $F(2, 61) = 7.66$	p<.001	p<.001	α=0.05	Large (.20)	✓

To probe the interaction effect, simple main effects tests were performed, which revealed a significant large main effect of Time for students in the feedback condition, [Wilks' Lambda = 0.84, $F(2,30) = 5.75$, $p = .005$, ηp2 = .16], and a significant moderate main effect of Time for students in the no-feedback condition, [Wilks' Lambda = 0.87, $F(2,30) = 4.55$, $p = .014$, ηp2 = .13].

Between Time1 and Time2, for students who received feedback, simple effects tests revealed that there was no significant difference in W/T, $t(31) = -1.83$, $p = .08$, $d = .32$. For students who did not receive feedback, simple effects tests revealed that there was a significant difference, $t(31) = 2.18$, $p = .04$, $d = .39$, indicating a small effect, such that mean number of W/T at Time2 was significantly lower (a decline) than Time1.

Between Time2 and Time3, for students who received feedback, simple effects tests revealed that there was a significant difference between mean number of W/T at Time2 and Time3, $t(31) = -2.16$, $p = .04$, $d = .38$, indicating a small effect, such that mean number of W/T at Time3 was significantly higher than Time2. For students who did not receive feedback, simple effects revealed that there was no significant difference, $t(31) = -0.68$, $p = .50$, $d = .12$.

Between Time1 and Time3, for students who received feedback, simple effects tests revealed that there was a significant difference, $t(31) = -3.49$, $p<.001$, $d = .62$, indicating a medium effect, such that mean number of W/T was significantly higher at Time3 than Time1. For students who did not receive feedback, simple effects revealed that there was a significant difference, $t(31) = 2.29$, $p = .03$, $d = .41$, indicating a small effect, such that mean number of W/T was significantly lower (a decline) at Time3 than Time1.

4.1.4.2 Words per Error-Free T-Unit (W/EFT)

Table 4-34 provides the descriptive statistics for W/EFT in students' writing across time as split out for feedback condition. The raw figures suggest that students in the no-feedback group started at a higher mean number of W/EFT at Time1 than the feedback group. A Mann-Whitney U test indicated that there was a statistically significant difference between the two

groups' mean number of W/EFT at Time1. [$U = 265.50$, $z = -3.316$, $p<.001$], with the no-feedback group having produced on average more W/EFT in their first performances. The raw figures suggest that the feedback group had more W/EFT with repetitions. The no-feedback group seemed to have fairly similar W/EFTs across repetitions (and they started off with more than the feedback group).

Table 4-34. *Descriptive statistics: W/EFT (by feedback condition)*

	Time1		Time2		Time3	
	Feedback Group	No-Feedback Group	Feedback Group	No-Feedback Group	Feedback Group	No-Feedback Group
Mean	9.79	13.12	11.48	13.84	15.10	12.98
SD	3.56	5.86	1.71	3.04	4.99	2.33
Minimum	0.00	0.00	8.00	9.17	9.22	9.57
Maximum	12.88	22.67	14.71	18.40	26.14	17.36

Figure 4.16 visualizes these findings for Time and Feedback group.

Figure 4.16. *Mean: W/EFT (by feedback condition)*

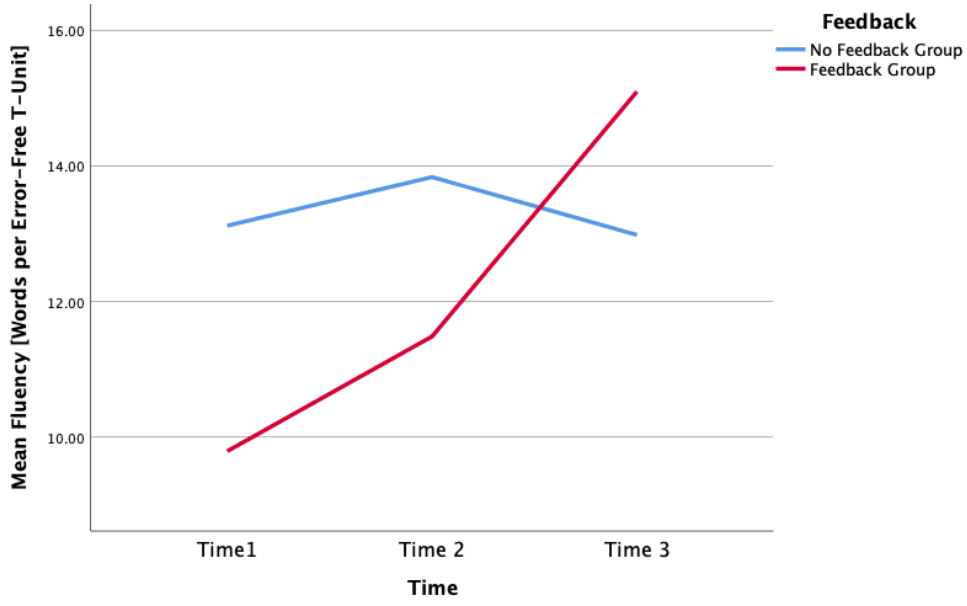


Table 4-35 provides the comparative statistics for mean number of W/EFT, including Holm-Bonferroni adjustments to control for Type 1 errors. A 2 (Feedback: Feedback v. No Feedback) x 3 (Time: Time1 v. Time2 v. Time3) between-within subjects Analysis of Variance (ANOVA) was conducted to assess the effects of Feedback and Time (Task Repetition) on W/EFT. This analysis revealed a significant, large interaction effect, [Wilks' Lambda = 0.69, $F(2, 61) = 13.8$, $p_{adjusted}<.001$, $\eta_p^2=.31$]. It also revealed a significant, large main effect of Time, [Wilks' Lambda = 0.78, $F(2, 61) = 8.9$, $p_{adjusted}<.001$, $\eta_p^2=.23$], i.e. the students wrote more

W/EFT on average after repetitions. However, there was no main effect of Feedback, [$F(1, 62) = 3.85$, $p_{adjusted} = .108$, $\eta_p^2 = .06$].

Table 4-35. *Comparative statistics: W/EFT*

	ANOVA Test	Unadjusted *p*	Holm Bonferroni adjusted *p*	Holm Bonferroni adjusted α	Effect Size (η_p^2)	Hypothesised Direction
Time	Wilks' Lambda =0.78 $F(2, 61) = 8.9$	$p<.001$	$p<.001$	α=.025	Large (.23)	✓
Feedback condition	$F(1, 62)=3.85$	Not significant $p=.054$	$p=.108$	α=.025		
Time x Feedback condition	Wilks' Lambda = 0.69 $F(2, 61) = 13.8$	$p<.001$	$p<.001$	α=.025	Large (.31)	✓

To probe the interaction effect, simple main effects tests were performed, which revealed a significant large main effect of Time for students in the feedback condition, [Wilks' Lambda = 0.58, $F(2,30) = 22.01$, $p<.001$, ηp2 = .42]. However, there was no effect of Time for students in the no-feedback condition, [Wilks' Lambda = 0.98, $F(2,30) = .67$, $p = .52$, ηp2 = .02].

Between Time1 and Time2, for students who received feedback, simple effects tests revealed that there was a significant difference, $t(31) = -2.22$, $p = .03$, $d = .39$, indicating a small effect, such that mean number of W/EFT at Time2 was significantly higher than Time1. For students who did not receive feedback, simple effects tests revealed that there was no significant difference, $t(31) = -0.56$, $p = .58$, $d = .10$.

Between Time2 and Time3, for students who received feedback, simple effects tests revealed that there was a significant difference between mean number of W/EFT, $t(31) = -4.04$, $p <.001$, $d = .71$, indicating a medium effect, such that mean number of W/EFT at Time3 was significantly higher than Time2. For students who did not receive feedback, simple effects revealed that there was no significant difference, $t(31) = 1.58$, $p = .12$, $d = .28$.

Between Time1 and Time3, for students who received feedback, simple effects tests revealed that there was a significant difference, $t(31) = -7.27$, $p<.001$, $d = 1.29$, indicating a large effect, such that mean number of W/EFT was significantly higher at Time3 than Time1. For students who did not receive feedback, simple effects revealed that there was no significant difference, $t(31) = .12$, $p = .90$, $d = .02$.

4.1.5 Conclusions on the effects of task repetition and feedback on CAF

Regarding the impact of repetition and feedback (RQ1), there were interaction effects of repetition and feedback on CAF in the written performances. The following three tables (Tables Table 4-36, Table 4-37, and Table *4-38*) collectively summarize the main results for RQ1. While Tables Table 4-36 and Table 4-37 highlight both significant and non-significant outcomes across the key variables, Table 4-38 provides a more fine-grained picture of the interaction effects and is essential for interpreting the observed trends in full. These three tables help show complementary perspectives of the interaction of repetition and feedback, thereby each table providing different layers of detail to help interpret the nuances of the results. It should be noted that the hypothesised directions, i.e., expected trends, varied slightly for methodological reasons. In Table 4-36, predicted directions (⇑/⇓) of the raw numerical data – always representing improvement in writing quality – are shown for all measures to display predictions.

Table 4-36 presents the full set of CAF measures for the interaction between repetition and feedback, along with measures that did not show a significant interaction effect. In doing so, the table distinguishes statistically significant from non-significant findings for each variable. It also compares and contrasts the hypothesised direction, i.e., the predicted direction of each measure's means, with the observed written performances, thereby identifying which measures aligned with or diverged from predicted patterns.

All *hypothesised directions* were set as predicted improvements that were shown by increases (⇑) of values for the measures or decreases (⇓), regardless of feedback condition, even though the actual observed outcomes diverged in some cases. For example, decreases (⇓) in errors per 100 words, errors per T-unit, or in the ratio of simple to complex sentences indicate improvement in writing quality. For the other measures, predicted increases (⇑) indicate improvements in writing quality.

The *observed performances* are categorized as significant increases, decreases, or no significant change, with symbols used for clarity: "↑" = statistically significant increase, "↓" = significant decrease, and "x" = no significant change. Decreases in errors/100 words, E/T, or in the ratio of simple to complex sentences are improvements in writing quality; decreases in errors for the other measures are declines in writing quality. Performance descriptions are also provided. The seven CAF measures showing significant interaction effects are highlighted in blue.

Table 4-36. *Interaction between repetition x feedback condition*

Measure	*p*	Effect size η_p^2	Hypothesised direction	Observed performance	Improved or declined writing quality
Complexity					
Average sentence length	*p*=.008	Large (.18)	⇑	FB↑; NFBx	FB: Improved; NFB: *Non-sig.*
Ratio: Simple/complex	*p*=.009	Large (.17)	⇓	FB↓; NFBx	FB: Improved; NFB: *Non-sig.*
Clauses per T-unit (C/T)	*p*=.495 *Non-sig.*	Small (.02)	⇑	x	*Non-sig.*
Lexical diversity	*p*=.158 *Non-sig.*	Moderate (.08)	⇑	↑	Improved
Sophisticated words	*p*<.001	Large (.29)	⇑	↑	Improved
Lexical sophistication	*p*<.001	Large (.37)	⇑	FB↑; NFB↓	FB: Improved; NFB: Declined
Accuracy (all errors)					
Errors per 100 words	*p*=.147 *Non-sig.*	Moderate (.09)	⇓	↓	Improved
E/T	*p*=.794 *Non-sig.*	Small (.02)	⇓	↓	Improved
EFT/T	*p*=.794 *Non-sig.*	Small (.03)	⇓	↓	Improved
Accuracy (by error type per 100 words)					
Subject-verb agreement	*p*=.698 *Non-sig.*	Small (.03)	⇓	↓	Improved
Verb tense	*p*=.452 *Non-sig.*	Moderate (.07)	⇓	x	*Non-sig.*
Verb form	*p*=.698 *Non-sig.*	Small (.02)	⇓	↓	Improved
Prepositions	*p*<.001	Large (.25)	⇓	↓	Improved
Articles	*p*=.452 *Non-sig.*	Moderate (.06)	⇓	x	*Non-sig.*
Fluency					
W/T	*p*<.001	Large (.20)	⇑	FB↑; NFB↓	FB: Improved; NFB: Declined
W/EFT	*p*<.001	Large (.31)	⇑	FB↑; NFBx	FB: Improved; NFB: *Non-sig.*

Notes:

1. "FB" = feedback group; "NFB" = no-feedback group
2. "⇑" = hypothesised increase in the measure; "⇓" = hypothesised decrease; "↑" without FB/NFB = significant increase for both groups; "↓" without FB/NFB = significant decrease for both groups
3. "Non-sig." = not significant
4. "x" = no statistically significant change (if no FB/NFB indications, then no significant changes for either group)
5. Decreases in errors/100 words, E/T, or in the ratio of simple to complex sentences indicate improvements; increases indicate declines.
6. Effect sizes are reported for both significant and non-significant p-values for fuller transparency of the data (Plonsky & Oswald, 2014).

7. Effect size values much greater than (.14) can legitimately occur in repeated-measures designs such as ANOVA (Lakens, 2013).

Table 4-37 shows the simple main effects of time split out for the two feedback groups in terms of the seven measures with interaction effects between repetition and feedback condition. In addition, this table identifies the general trends for the data, i.e., the effect of time/repetition on the performance of each of the two groups. It is important to note that descriptors used in this table, such as "slight" (very small changes), "somewhat" (modest or moderate changes), "plateau" (leveling off of performance after initial change), "linear" (roughly straight-line progression over time), and "non-linear" (variation rather than steady progression) are used qualitatively to characterize patterns observed in the respective descriptive figures in section 4.1. They are intended only as interpretive aids, i.e., qualitative visual interpretation of the figures, to approximate characterizations of observed visual trends. Therefore, they should not be treated as precise quantitative categories or statistical classifications.

Table 4-37. *Simple main effects of Time [measures with interaction effects]*

Measure	Corresponding figure	Simple effects of Time					
		Feedback group			No-feedback group		
		p	Effect size η_p^2	Trend	*p*	Effect size η_p^2	Trend
Complexity							
Average sentence length	Figure 4.1	*p*=.002	Large (.34)	Linear: steady improvement over time	*p*=.009	Large (.27)	Non-linear: initial decline followed by partial recovery toward baseline
Ratio: Simple/complex	Figure 4.2	*p*=.003	Large (.33)	Linear: steady improvement over time	*p*=.124 *Non-sig.*	Moderate (.13)	Non-linear: decline then recovery to baseline (non-sig.)
Sophisticated words	Figure 4.5	*p*<.001	Large (.65)	Non-linear: initial improvement followed by a steadier upward trend	*p*<.001	Large (.52)	Non-linear: improve-ment followed by a pla-teau
Lexical sophistication	Figure *4.6*	*p*<.001	Large (.25)	Non-linear: slight decline before	*p*<.001	Large (.21)	Non-linear: initial improve-

				recovering to higher level				ment followed by decline
Accuracy (by error type per 100 words)								
Prepositions	Figure *4.13*	$p<.001$	Large (.66)	Non-linear: very steady improvement then a somewhat steady improvement		$p<.001$	Large (.50)	Non-linear: decline followed by slight improve-ment
Fluency								
W/T	Figure *4.15*	$p=.005$	Large (.16)	Non-linear: somewhat initial improvement followed by a steadier upward trend		$p=.014$	Moderate (.13)	Non-linear: initial decline then somewhat of a plateau
W/EFT	Figure *4.16*	$p<.001$	Large (.42)	Non-linear: somewhat initial improvement followed by a steadier upward trend		$p=.52$ *Non-sig.*	Small (.02)	Non-linear: slight improve-ment then a decline back to baseline (non-sig.)

Notes:

1. The seven measures that had an interaction effect are colored in blue.
2. "Non-sig." = not significant
3. "Linear" = describes data that follow a straight line
4. "Non-linear" = describes data that do not follow a straight line
5. Decreases in errors/100 words, E/T, or in the ratio of simple to complex sentences are improvements; increases are declines.
6. "Slight", "somewhat", and "plateau" are not statistical categories but simply visual interpretations
7. "Increase/decrease" refers to numerical change; "improve/decline" refers to qualitative outcome

Going a step further, Table 4-38 presents data that show the statistically significant improvements and declines, as well as non-significant differences, based on time periods as split out by feedback condition for all CAF measures. These data also suggest that there were some trade-off effects for CAF. As introduced in section 2.1, the trade-off hypothesis predicts that when students perform a task for the first time, their cognitive capacities are limited such that they focus their concentration on applying their skills to complete the task (Skehan, 1998a). As one is completing the task, the CAF dimensions compete with one another, very frequently between accuracy and complexity, as the student learns the demands of the task completion.

This table shows the measures where the differences in mean scores were statistically significant, improved [shown in blue] or declined [shown in red], from Time1 to Time2, Time2 to Time3, and Time1 to Time3, as well as measures with mean differences that were not

statistically significant [shown in black]. The seven CAF measures with an interaction effect are listed in blue.

Table 4-38. *Written performances: Improvements/Declines/Similarities (by time and feedback condition)*

	Feedback Group				No-Feedback Group		
Measure	Time 1-2	Time 2-3	Time 1-3		Time 1-2	Time 2-3	Time 1-3
Complexity							
Average sentence length	*Non-sig.*	*Non-sig.*	Improved		Declined	Improved	*Non-sig.*
Ratio: Simple/complex	Improved	*Non-sig.*	Improved		*Non-sig.*	*Non-sig.*	*Non-sig.*
Clauses per T-unit (C/T)	*Non-sig.*	*Non-sig.*	*Non-sig.*		*Non-sig.*	*Non-sig.*	*Non-sig.*
Lexical diversity	*Non-sig.*	Improved	Improved		*Non-sig.*	Improved	Improved
Sophisticated words	Improved	Improved	Improved		Improved	*Non-sig.*	Improved
Lexical sophistication	*Non-sig.*	Improved	Improved		*Non-sig.*	Declined	Declined
Accuracy							
Errors/100 words	Improved	Improved	Improved		Improved	Improved	Improved
Errors per T-unit (E/T)	Improved	Improved	Improved		Improved	Improved	Improved
Error-free T-units/T-unit (EFT/T)	*Non-sig.*	Improved	Improved		*Non-sig.*	Improved	Improved
Subject-verb errors/100 words	*Non-sig.*	Improved	Improved		*Non-sig.*	Improved	Improved
Verb tense errors/100 words (VT)	*Non-sig.*	*Non-sig.*	*Non-sig.*		*Non-sig.*	*Non-sig.*	*Non-sig.*
Verb form errors/100 words (VF)	*Non-sig.*	Improved	Improved		*Non-sig.*	Improved	Improved
Preposition errors/100 words	Improved	Improved	Improved		*Non-sig.*	Improved	Improved
Article errors/100 words	*Non-sig.*	*Non-sig.*	*Non-sig.*		*Non-sig.*	*Non-sig.*	*Non-sig.*
Fluency							
Words per T-unit (W/T)	*Non-sig.*	Improved	Improved		Declined	*Non-sig.*	Declined
Words per error-free T-unit (W/EFT)	Improved	Improved	Improved		*Non-sig.*	*Non-sig.*	*Non-sig.*

Notes:

1. Decreases in errors/100 words, E/T, or in ratio of simple to complex sentences are improvements; increases are declines.

Color codes:

2. Blue = scores have significantly improved between the specified times, for example, Time 1-Time 2.
3. Red = scores have significantly declined
4. Black = no significant differences in the scores

Overall, both groups showed some improvements on several measures over time, with more improvements in the feedback group. However, improvements did not occur at every time interval or on every measure**,** and some measures were stable or declined. For the feedback group, there were significant improvements in some CAF measures at each time. Also, for the feedback group, there were significant improvements in some CAF measures, but at the expense of some other CAF measures where there were no significant differences rather than significant declines, i.e., a trade-off where some measures significantly improved at the expense of (or in competition with) other CAF measures where there were non-significant differences, as opposed to significant gains occurring simultaneously with significant declines. As introduced in section 2.1, a trade-off does not require significant performance gains to be automatically accompanied by a significant loss for there to be a trade-off; it can also simply be accompanied by a similar level of performance rather than a gain. For the no-feedback

group, there were significant improvements in accuracy and some specific complexity measures, with significant declines in some complexity and fluency measures, i.e., a trade-off where some measures significantly improved at the expense of (or in competition with) other CAF measures where there were significant declines, as well as some measures where there were non-significant differences. The trade-offs that occurred are identified and discussed in section 5.2.1.

Together, the findings across Tables Table *4-36*, Table *4-37*, and Table 4-38 offer an integrated perspective of ways that time/repetition and feedback interact to influence CAF measures in student writing. Table 4-36 identifies specific measures where there were statistically significant interaction effects, while showing differences in performance directions between the two groups. Table 4-37 builds on the information from Table 4-36 for the seven CAF measures where there was an interaction effect by breaking down the interaction effects for each group over time while revealing performance trends that suggest linear and non-linear patterns as well as plateaued trends, thus providing additional depth for interpreting the findings. Table 4-38 complements the insights from Tables Table *4-36* and Table *4-37* by displaying an overview of the statistically significant improvements and declines, as well as non-significant changes for all CAF measures and time intervals, i.e., Time1 to Time2, etc., separated by the feedback and no-feedback groups. This fine-grained synthesis from Table 4-38 is consistent with the trends observed in Tables Table *4-36* and Table *4-37*, suggesting that the feedback group shows more consistent gains in CAF while the no-feedback group shows mixed outcomes, e.g., some declines, etc. This integrated overview provides some support for the trade-off hypothesis by revealing where trade-offs occur.

In terms of repetition (RQ2), the results in terms of CAF measures reported in section 4.1 show that students' EAP writing elicited by the listening-to-write task employed in this study statistically significantly improved with repetitions, as based on several measures. Table 4-39 displays a detailed overview of the effects of repetition on CAF performance indicators for which this was the case, i.e., a significant main effect of repetition, as well as for those for which it was not the case. As shown in Table 4-39, the directional arrows reflect the direction of numerical values. For errors per 100 words, errors per T-unit and the ratio of simple to complex sentences, decreases (↓) represent improvements; for the other measures, increases (↑) represent improvements; and "x" represents no significant change. This clarification ensures consistency with the patterns shown in Table 4-38. While both groups made improvements over time on various variables, a greater number of improvements were observed in the

feedback group. Findings for the no-feedback group suggest more mixed outcomes, such as some significant declines or non-significant changes.

All hypothesised directions were set as predicted improvements shown by increases (⇑) or decreases (⇓) of values for the measures. In terms of results that aligned with the hypothesised improvement predictions, this table shows (✓) only when a statistically significant main effect confirmed the predicted improvement. For non-significant main effects, no alignment indicator is shown, and (✓) therefore indicates improvement supported by the overall main effect, not by isolated time interval comparisons, e.g., T1–T2, etc. The nine CAF measures that had a main effect of repetition are listed in blue.

Table 4-39. *Effect of repetition: CAF*

Measure	*p*	Effect size η_p^2	Hypothesised direction	Result alignment with hypothesised direction	Significant increases or decreases		
					T1 →T2	T2 →T3	T1 →T3
Complexity							
Average sentence length	*p*=.084 *Non-sig.*	Moderate (.12)	⇑		FBx; NFB↓	FBx; NFB↑	FB↑; NFBx
Ratio: Simple/complex	*p*=.30 *Non-sig.*	Moderate (.06)	⇓		FB↓; NFBx	x	FB↓; NFBx
Clauses per T-unit (C/T)	*p*=.81 *Non-sig.*	Small (.01)	⇑		x	x	x
Lexical diversity	*p*<.001	Large (.23)	⇑	✓	x	↑	↑
Sophisticated words	*p*<.001	Large (.58)	⇑	✓	↑	FB↑; NFBx	↑
Lexical sophistication	*p*=.705 *Non-sig.*	Small (.01)	⇑		x	FB↑; NFB↓	FB↑; NFB↓
Accuracy (all errors)							
Errors per 100 words	*p*<.001	Large (.76)	⇓	✓	↓	↓	↓
E/T	*p*<.001	Large (.60)	⇓	✓	↓	↓	↓
EFT/T	*p*<.001	Large (.34)	⇓	✓	x	↑	↑
Accuracy (by error type per 100 words)							
Subject-verb agreement	*p*<.001	Large (.52)	⇓	✓	x	↓	↓
Verb tense	*p*=.27 *Non-sig.*	Small (.04)	⇓		x	x	x
Verb form	*p*<.001	Large (.22)	⇓	✓	x	↓	↓

Prepositions	*p*<.001	Large (.55)	⇓	✓	FB↓; NFBx	↓	↓
Articles	*p*=.088 *Non-sig.*	Moderate (.10)	⇓		x	x	x
Fluency							
W/T	*p*=.08 *Non-sig.*	Moderate (.08)	⇑		FBx; NFB↓	FB↑; NFBx	FB↑; NFB↓
W/EFT	*p*<.001	Large (.23)	⇑	FB✓; NFBx	FB↑; NFBx	FB↑; NFBx	FB↑; NFBx

Notes:

1. "FB" = feedback group; "NFB" = no-feedback group
2. "⇑" = hypothesised increase in the measure; "⇓" = hypothesised decrease; "↑" without FB/NFB = significant increase for both groups; "↓" without FB/NFB = significant decrease for both groups
3. "Non-sig." = not significant
4. "✓" = predicted improvement with repetition statistically confirmed by the main effect
5. "x" = no statistically significant change (if no FB/NFB indications, then no significant changes for either group)
6. Decreases in errors/100 words, E/T, or in the ratio of simple to complex sentences indicate improvements; increases indicate declines.
7. Effect sizes are reported for both significant and non-significant p-values for fuller transparency of the data (Plonsky & Oswald, 2014).
8. Effect size values much greater than (.14) can legitimately occur in repeated-measures designs such as ANOVA (Lakens, 2013).

It is further worthwhile noting that the two feedback-condition groups did not always have similar means for the CAF measures on their writing performances at Time1. Table 4-40 shows that there were statistically significant differences for six of the CAF measures in the two-group comparison at Time1, as calculated with Mann-Whitney U tests (given the non-normal distributions of many measures). With the exception of two measures (Ratio simple to complex sentences and preposition errors per 100 words), the measure was higher in those cases for the no-feedback group. It is additionally important to note that the students had been allocated randomly to feedback conditions, and the overall student sample had been controlled as much as possible for L2 characteristics, as introduced earlier in sections 3.2 and 3.4. Thus, there is no clear reason available for why some of these differences were observed. Regardless, these differences at Time1 did not systematically associate with effects of repetition or feedback, and their interaction, as for some of these measures' effects were observed and for others not.

Table 4-40. *Feedback group comparison at Time1 (CAF)*

Measure	*p*	Similarity at Time1 (✓ or X)	Result FB vs. no-FB	Explanation
Complexity				
Average sentence length	*p* = .029	X	FB < noFB	FB: Worse
Ratio: Simple sentence to complex sentence	*p* = .041	X	FB > noFB	FB: Worse
C/T	*p* = .019	X	FB < noFB	FB: Worse

Lexical diversity	$p = .186$	✓		
Sophisticated words	$p = .007$	X	FB < noFB	FB: Worse
Lexical sophistication	$p = .203$	✓		
Accuracy (by global measure)				
Errors per 100 words	$p = .075$	✓		
E/T	$p = .835$	✓		
EFT/T	$p = .291$	✓		
Accuracy (by error type)				
Subject-verb agreement per 100 words	$p = .941$	✓		
Verb tense per 100 words	$p = .143$	✓		
Verb form per 100 words	$p = .952$	✓		
Preposition errors per 100 words	$p > .001$	X	FB > noFB	FB: Worse
Article errors per 100 words	$p = .921$	✓		
Fluency				
W/T	$p = .162$	✓		
W/EFT	$p < .001$	X	FB < noFB	FB: Worse

Note: "FB" = feedback group; "NoFB" = no-feedback group

In terms of *feedback* (RQ3), there was no main effect of feedback on students' academic writing performances in terms of CAF.

4.2 Knowledge summary and transfer

Table 4-41 summarizes the measures used to establish the extent to which the students were able to summarize the knowledge from the listening input into their writing and transfer the input content to a new context as demonstrated in their writing. As shown in **Error! Reference source not found.**, this was measured with a modified TOEFL integrated task rubric and an adapted AACU rubric as introduced in section 3.5.3.2.

Table 4-41. *Knowledge summary and transfer measures*

Construct	Measure	Evidence of improvement	Analysis
Knowledge summary	TOEFL integrated task rubric (modified)	Increased holistic score	Manual rating
Knowledge transfer	AACU rubric (modified)	Increased holistic score	Manual rating

4.2.1 Knowledge summary

Table 4-42 provides the descriptive statistics for knowledge summary in students' writing across time as split out for feedback condition. The raw figures suggest that both groups started at very similar means for knowledge summary at Time1. A Mann-Whitney U test indicated that there was no statistically significant difference between the two groups' mean knowledge

summary score at Time1. [$U = 416.50$, $z = -1.341$, $p = .180$], with both groups having scored on average quite similarly on knowledge summary in their first performances. The raw figures suggest that both groups follow the pattern of continuing to improve their written summary to very similar extents with repetitions. The maximum mean score that could be reached was 5.

Table 4-42. *Descriptive statistics: Knowledge summary (by feedback condition)*

	Time1		Time2		Time3	
	Feedback Group	No-Feedback Group	Feedback Group	No-Feedback Group	Feedback Group	No-Feedback Group
Mean	3.09	3.06	3.66	3.72	4.28	4.28
SD	1.03	0.95	1.10	0.89	0.89	0.77
Minimum	1.00	2.00	2.00	2.00	2.00	3.00
Maximum	5.00	4.00	5.00	5.00	5.00	5.00

Figure 4.17 visualizes these findings for Time and Feedback group.

Figure 4.17. *Mean: Knowledge summary (by feedback condition)*

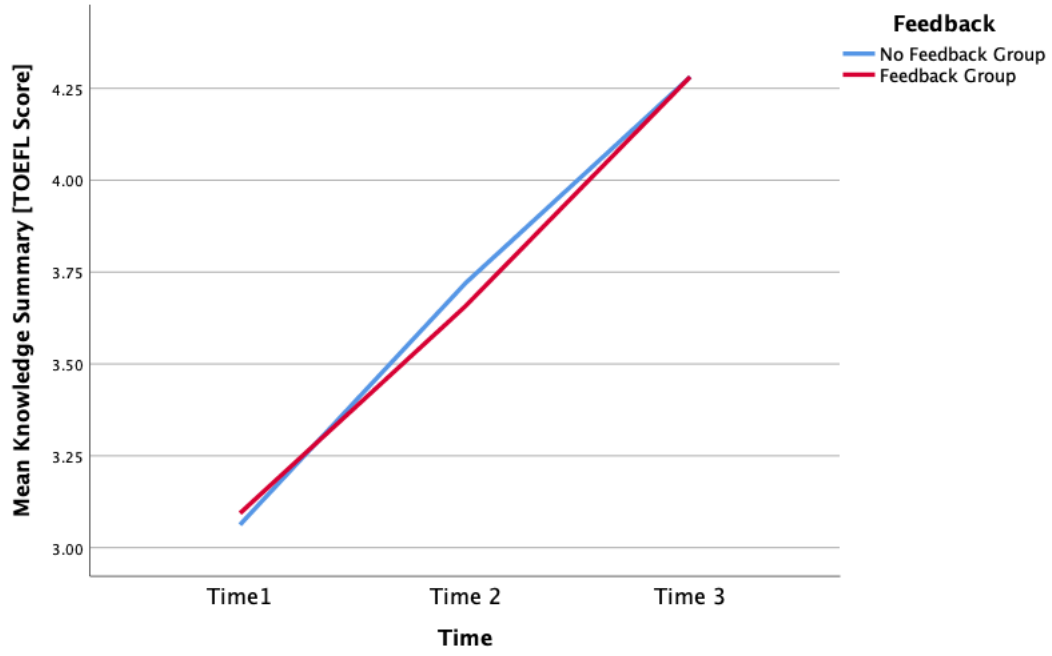


Table 4-43 provides the comparative statistics for average knowledge summary scores. A 2 (Feedback: Feedback v. No Feedback) x 3 (Time: Time1 v. Time2 v. Time3) between-within subjects Analysis of Variance (ANOVA) was conducted to assess the effects of Feedback and Time (Task Repetition) on average knowledge summary scores. This analysis revealed no significant interaction effect, [Wilks' Lambda = 1.0, $F(2, 61) = 0.11$, $p = .896$, $\eta_p^2 = .004$], as well as no significant main effect of Feedback, [$F(1, 62) = .003$, $p = .96$, $\eta_p^2 = .000$]. However, this analysis revealed a large main effect of Time, [Wilks' Lambda = 0.28,

$F(2, 61) = 78.96$, $p<.001$, $\eta_p^2 = .72$], i.e., the students made improvements on average knowledge summary scores after repetitions.

Table 4-43. *Comparative statistics: Knowledge summary*

	ANOVA Test	Statistical Significance	Effect Size (η_p^2)	Hypothesised Direction
Time	Wilks' Lambda = 0.28 $F(2, 61) = 78.96$	$p<.001$	Large (.72)	✓
Feedback condition	$F(1, 62)=.003$	Not significant $p=0.96$		
Time x Feedback condition	Wilks' Lambda = 1.0 $F(2, 61) = 0.11$	Not significant $p=.896$		

Simple effects tests revealed that there was a significant difference in average knowledge summary scores between Time1 and Time2, $t(63) = -6.18$, $p<.001$, $d = .77$, a significant difference between Time2 and Time3, $t(63) = -5.47$, $p<.001$, $d = .68$, as well as a significant difference between Time1 and Time3, $t(63)= -12.7$, $p<.001$, $d = 1.58$.

4.2.2 Knowledge transfer

Table 4-44 provides the descriptive statistics for knowledge transfer in students' writing across time as split out for feedback condition. The raw figures suggest that both groups started at very similar means for knowledge transfer at Time1. A Mann-Whitney U test indicated that there was no statistically significant difference between the two groups' mean knowledge summary score at Time1. [$U = 456.00$, $z = -.842$, $p = .400$], with both groups having scored on average quite similarly on knowledge transfer in their first performances. The raw figures suggest that both groups continued to improve their ability to demonstrate knowledge transfer through their writing in somewhat similar amounts with repetitions. The maximum mean score that could be reached was 4.

Table 4-44. *Descriptive statistics: Knowledge transfer (by feedback condition)*

	Time1		Time2		Time3	
	Feedback Group	No-feedback Group	Feedback Group	No-feedback Group	Feedback Group	No-feedback Group
Mean	2.25	2.31	2.44	2.69	3.56	3.53
SD	0.76	0.47	0.80	0.97	0.80	0.67
Minimum	1.00	2.00	1.00	1.00	2.00	2.00
Maximum	3.00	3.00	3.00	4.00	5.00	4.00

Figure 4.18 visualizes these findings for Time and Feedback group.

Figure 4.18. *Mean: Knowledge transfer (by feedback condition)*

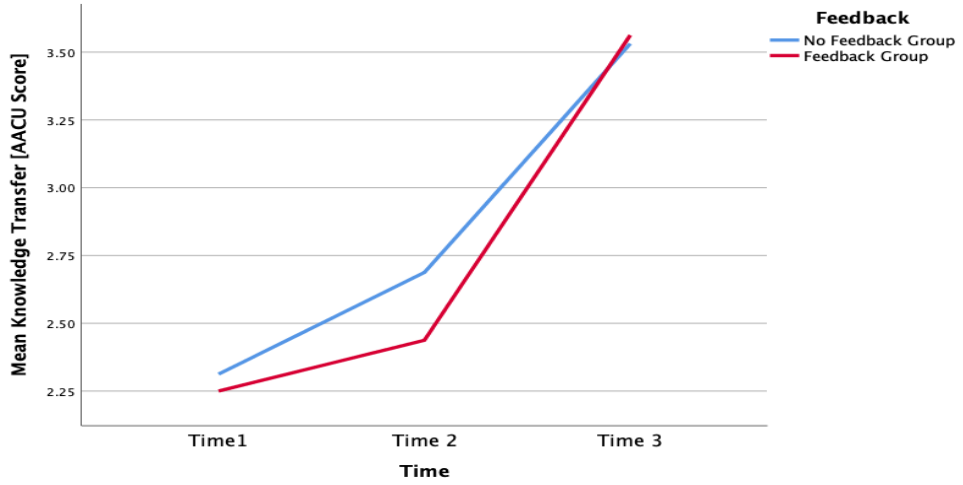


Table 4-45 provides the comparative statistics for average knowledge transfer scores. A 2 (Feedback: Feedback v. No Feedback) x 3 (Time: Time1 v. Time2 v. Time3) between-within subjects Analysis of Variance (ANOVA) was conducted to assess the effects of Feedback and Time (Task Repetition) on average knowledge transfer scores. This analysis revealed no significant interaction effect, [Wilks' Lambda = 0.95, $F(2, 61) = 1.64$, $p = .20$, $\eta_p^2 = .05$], as well as no significant main effect of Feedback, [$F(1, 62) = .31$, $p = .577$, $\eta_p^2 = .01$]. However, this analysis revealed a large main effect of Time, [Wilks' Lambda = 0.20, $F(2, 61) = 120.71$, $p<.001$, $\eta_p^2 = .80$], i.e., the students made improvements on average knowledge transfer scores after repetitions.

Table 4-45. *Comparative statistics: Knowledge transfer*

	ANOVA Test	Statistical Significance	Effect Size (η_p^2)	Hypothesised Direction
Time	Wilks' Lambda=0.20 $F(2, 61) = 120.71$	$p<.001$	Large (.80)	✓
Feedback condition	$F(1, 62)=0.31$	Not significant $p=.577$		
Time x Feedback condition	Wilks' Lambda = 0.95 $F(2, 61) = 1.64$	Not significant $p=0.20$		

Simple effects tests revealed that there was a significant difference in average knowledge summary scores between Time1 and Time2, $t(63) = -4.1$, $p<.001$, $d = .51$, a significant difference between Time2 and Time3, $t(63) = -11.6$, $p<.001$, $d = 1.45$, as well as a significant difference between Time1 and Time3, $t(63)= -15.6$, $p<.001$, $d = 1.95$.

4.2.3 Conclusions on the effects of task repetition and feedback on knowledge summary & transfer

In terms of the *effect of task repetition and feedback* (RQ1), there were no interaction effects of repetition and feedback for knowledge summary and transfer. In terms of *feedback* alone, there was no main effect. Moreover, the accompanying effect sizes were uniformly small. This suggests that feedback was not likely to have substantially influenced knowledge summary or transfer in this study. Having established that interaction effects (RQ1) and feedback alone (RQ3) did not show substantial influence, I now turn to task repetition (RQ2) where the significant and most notable changes occurred for knowledge summary and transfer. In terms of *repetition* (RQ2), however, students' EAP writing elicited by the listening-to-write task employed in this study statistically significantly improved with repetitions, with large effect sizes for both indicators. The hypothesised directions were set as predicted improvements shown by increases (⇑) of values for the measures. The directional arrows (↑) reflect the direction of numerical values. As Table 4-46 shows, both indicators aligned with the hypothesised improvement predictions (✓), with statistically significant main effects confirming these predicted improvements. This table also shows where there were significant improvements between repetitions (T1-T2, etc.).

Table 4-46. *Effect of repetition: Knowledge summary and transfer*

Measure	*p*	Effect size η_p^2	Hypothesised direction	Result alignment with hypothesised direction	Significant improvements		
					T1→T2	T2→T3	T1→T3
Knowledge summary and transfer							
Knowledge summary	$p<.001$	Large (.72)	⇑	✓	↑	↑	↑
Knowledge transfer	$p<.001$	Large (.80)	⇑	✓	↑	↑	↑

Notes:

1. "⇑" = hypothesised increase of value of the measure
2. "↑" = statistically significant increase
3. "↑" without FB/NFB indications = significant increases for both groups
4. "✓" = predicted improvement with repetition statistically confirmed by the main effect

It is worth noting that the two feedback-condition groups had similar means for the measures for knowledge summary and transfer on their writing performances at Time1. Table *4-47* shows that there were no statistically significant differences for the two measures in the two-group comparison at Time1, as calculated with Mann-Whitney U tests.

Table 4-47. *Feedback group comparison at Time 1 (Knowledge summary and transfer)*

Measure	*p*	Similarity at Time1

		(✓ or X)
Knowledge summary and transfer		
Knowledge summary	$p = 0.18$	✓
Knowledge transfer	$p = 0.40$	✓

4.3 Student perceptions of task repetition

In this section, I report the results on student perceptions of the task used in this study, of listening-to-write tasks more generally, and of the extent to which integrated task repetition helps develop academic writing proficiency (RQ4a). Then I report on the extent to which perceptions of task repetition differ between students who received feedback on their academic writing performance and those who did not (RQ4b).

To be able to answer these questions, student perceptions of integrated task repetition were investigated by means of responses to a post-task repetition perception questionnaire. Two categories of questions were posed and examined: Student perceptions about the listening-to-write task used in the study and about task repetition in general. Descriptive statistics were calculated for each category of questions. Then, comparative statistics were run to determine whether students' perceptions of the task and of task repetition were significantly different depending on whether they had received feedback.

4.3.1 Student perceptions about this task

In terms of students' perceptions about *this task* (RQ4a), Table 4-48 provides the perception statements, mean, standard deviation and frequency distributions for students' responses to the statements. Because not all statements were formulated in the same direction, I inversed the scale for the asterisked statements. So this means that the closer the mean is to 5, the more positive the students' perceptions were.

For all statements, the means were above 3.5 on the 5-point scale, which suggests that student views on the task were favorable. So, the closer to 5 that the means are, the more positive their perceptions were. Some examples of statements where more than half the participants agreed or strongly agreed (and others who selected "neutral") include "I enjoyed repeating the listening-to-write task", " The task accurately reflects my English listening ability", "The task reflects my English writing ability", "I had enough time to complete the writing task", and "the audio recording was [not] too long." However, there was a bit more variation in levels of agreement in terms of whether repeating this task was boring in comparison to the other statements: While the majority (60%) did not think it was boring, a

few students (18%) felt it was (and 22% were neutral about it). Overall, most students held positive views on the task.

Table 4-48. *Descriptive statistics: Student perceptions about this task*

Statement	*N*	*f*(%)					*M*(SD)
		1	2	3	4	5	
I enjoyed repeating the listening-to-write task	64	0(0%)	9(14%)	17(27%)	30(47%)	5(13%)	3.58(.89)
Repeating the listening-to-write task was boring*	64	6(9%)	6(9%)	14(22%)	17(27%)	21(33%)	3.64*(1.29)
The task reflects my English writing ability	64	0(0%)	3(5%)	20(31%)	30(47%)	11(17%)	3.77(.79)
The task accurately reflects my English listening ability	64	0(0%)	3(5%)	11(17%)	39(61%)	11(17%)	3.91(.73)
I had enough time to complete the writing task	64	3(5%)	0(0%)	9(14%)	43(67%)	9(14%)	3.86(.83)
The audio recording was played enough times for me to understand it	64	3(5%)	0(0%)	3(5%)	32(50%)	26(41%)	4.22(.92)
The audio recording provided sufficient ideas for me to complete the writing task	64	0(0%)	3(5%)	6(9%)	26(41%)	29(45%)	4.27(.82)
The audio recording was too long*	64	0(0%)	0(0%)	3(5%)	20(31%)	41(64%)	4.59*(.58)
Vocabulary in the audio recording was difficult for me*	64	0(0%)	3(5%)	21(33%)	28(44%)	12(19%)	3.77*(.81)
Sentence structures in the audio recording were complicated for me*	64	0(0%)	6(9%)	21(33%)	23(36%)	14(22%)	3.70*(.92)

Note: Student levels of agreement were selected on a Likert scale:
Strongly disagree (1); Disagree (2); Neutral (3); Agree (4); Strongly agree (5)
* = reverse coded variable

Table 4-49 provides the descriptive statistics for the total scores on the task perception questions for all participants together (RQ4a), as well as per feedback-condition group (RQ4b). This confirmed students' positive views overall (M=39.30 out of a total possible of 50).

Table 4-49. *Student perceptions about this task (by feedback group)*

Group	N	Mean	SD	Minimum	Maximum
Feedback	32	39.63	4.61	26	46
No-Feedback	32	38.97	4.68	26	46
All	64	39.30	4.62	26	46

To establish whether there was a significant difference between the two feedback condition groups' perceptions of the task, I first conducted a Kolmogorov-Smirnov test for normality. The task perception scores for both the Feedback group (p = .041) and No-Feedback group (p = .005) were not normally distributed. Therefore, a non-parametric Mann-Whitney U test was used to assess differences between the groups. This test indicated that there was no statistically significant difference between the two groups' overall task perceptions [U (N No-Feedback = 32, N Feedback = 32) = 477.50, z = -.466, p = .642].

4.3.2 Student perceptions about integrated listening-to-write tasks and task repetition

First, in terms of students' perceptions about *integrated listening-to-write tasks more generally* (RQ4a), Table 4-50 provides the statements, mean, standard deviation and frequency distributions for students' responses. For both statements "Writing after listening improves my writing" and "Listening with the purpose of writing helps me improve my English", the mean was high (4.45 and 4.55, respectively), and student responses generally ranged from agree (4) to strongly agree (5), with only a couple of "neutral" selections and no disagreement selections. Thus, students' views on integrated tasks were favorable. In sum, the students held positive views on the extent to which listening-to-write tasks help develop their writing and improve their English language proficiency.

Table 4-50. *Descriptive statistics: Student perceptions about listening-to-write tasks*

Statement	*N*	*f*(%)					*M*(SD)
		1	2	3	4	5	
Writing after listening improves my writing	64	0(0%)	0(0%)	3(5%)	29(45%)	32(50%)	4.45(.56)
Listening with the purpose of writing helps me improve my English	64	0(0%)	0(0%)	0(0%)	29(45%)	35(55%)	4.55(.50)

Note: Student levels of agreement were selected on a Likert scale:
Strongly disagree (1); Disagree (2); Neutral (3); Agree (4); Strongly agree (5)

Table 4-51 provides the descriptive statistics for the total scores on the perception questions on integrated listening-to-write tasks for all participants together (RQ4a), as well as per feedback condition group (RQ4b). This confirmed students' positive views overall about the way in which the integrated listening-to-write may help improve their writing (M=9.00 out of a total possible of 10).

Table 4-51. *Student perceptions about integrated listening-to-write tasks (by feedback group)*

Group	N	Mean	SD	Minimum	Maximum
Feedback	32	8.97	0.99	7	10
No-Feedback	32	9.03	1.03	7	10
All	64	9.00	1.01	7	10

To check whether perceptions differed between those who had and those who had not received feedback, comparative statistics were run. A Kolmogorov-Smirnov test for normality revealed that the task perception scores for both the Feedback group ($p<.001$) and No-Feedback group ($p<.001$) were not normally distributed. Therefore, a non-parametric Mann-Whitney U test was used to assess differences between the groups. This test indicated that there was no statistically significant difference between the feedback groups' task perceptions [U (N No-Feedback = 32, N Feedback = 32) = 493.00, z = -.275, p = .783].

Second, in terms of students' perceptions about *task repetition in general* (RQ4a), Table 4-52 provides the task repetition perception statements, mean, standard deviation and frequency distributions for students' responses. For all of the statements, the mean was above 4.0, and student responses generally ranged from agree (4) to strongly agree (5), which suggests that students' views on repeating tasks were favorable. Statements where all the participants agreed or strongly agreed were: "Repeating a task helps me improve my listening" and "Fulfilling a listening-to-write task gets easier with repetition." For the statements "Repeating a task helps me improve my writing" and "I would like to do task repetition in future classes", only 5% and 9% respectively responded to be neutral about it, and no one disagreed with these statements. In sum, the students held positive views on repeating tasks.

Table 4-52. *Descriptive statistics: Student perceptions about task repetition*

Statement	*N*	*f*(%)					*M*(SD)
		1	2	3	4	5	
Repeating a task helps me improve my writing	64	0(0%)	0(0%)	3(5%)	38(59%)	23(36%)	4.31(.56)

Repeating a task helps me improve my listening	64	0(0%)	0(0%)	0(0%)	29(45%)	35(55%)	4.55(.50)
Fulfilling a listening-to-write task gets easier with repetition	64	0(0%)	0(0%)	0(0%)	20(31%)	44(69%)	4.69(.47)
I would like to do task repetition in future classes	64	0(0%)	0(0%)	6(9%)	30(47%)	28(44%)	4.34(.65)

Note: Student levels of agreement were selected on a Likert scale:
Strongly disagree (1); Disagree (2); Neutral (3); Agree (4); Strongly agree (5)

Table 4-53 provides descriptive statistics for the total scores on the task repetition perception questions, for all participants together (RQ4a), as well as per feedback condition group (RQ4b). This confirms that students' views are overall positive (M=17.89 out of a total possible of 20), which is also the case in both feedback-condition groups.

Table 4-53. *Student perceptions about task repetition (by feedback group)*

Group	N	Mean	SD	Minimum	Maximum
Feedback	32	17.84	1.32	15	20
No-Feedback	32	17.94	1.46	15	20
All	64	17.89	1.38	15	20

To check whether perceptions differed between those who had and those who hadn't received feedback, comparative statistics were run. A Kolmogorov-Smirnov test for normality revealed that the task repetition perception scores for both the Feedback group (p = .002) and No-Feedback group (p<.001) were not normally distributed. Therefore, a non-parametric Mann-Whitney U test was used to assess differences between the groups. As expected on the basis of the means, there was no statistically significant difference between the two groups' general perceptions about task repetition [U (N No-Feedback = 32, N Feedback = 32) = 481.00, z = -.429, p=.668].

Third, in terms of the open-ended question *"What is your overall opinion about task repetition? Explain"* which also addresses (RQ4b), out of the 64 students who participated in this study, 59 (92.2%) responded to this question (28 responses from the Feedback group; 31 responses from the No-feedback group). Out of the 59 responses to this question, 51 (86.4%) were positive opinions (23 from the Feedback group, 28 from the No-Feedback group); 5 (8.5%) were negative opinions (2 from the Feedback group, 3 from the No-Feedback group); and 3 (5.1%) were mixed opinions (3 from the Feedback group, 0 from the No-Feedback

group). Out of the 5 students who did not respond to this question, 4 were from the Feedback group and 1 was from the No-Feedback group.

It is important to note that most of the students' opinions about task repetition included an explanation, i.e., some elaboration on what they found good about it or what they thought it might be good for. For example, within the majority of the positive opinions, some replies focused, for example, on being good for their listening while other replies focused on helping their writing development or on being positive for both their listening and writing skills. Some statements focused on improving their language skills more generally without specifying a special skill. Next, I delve into the various categories that I provided for the positive responses.

Regarding the *positive opinions*, out of the 51 total positive responses, 39 (76.5%) opinions included explanations. Based on the types of explanations that accompanied the positive opinions, I differentiated them into the following categories: positive mention of listening skills, positive mention of writing skills, positive mention of both listening and writing skills, and positive mention of general skills. Next are tables where I provide examples of student opinions for the above-mentioned categories.

There were a total of 12 (23.5%) positive mentions of listening out of the 51 total positive responses (4 from the Feedback group, 8 from the No-Feedback group). Table 4-54 provides examples of positive opinions that reflect the listening skill.

Table 4-54. *Examples of positive mention of listening*

Participant's comment
"Task repetition always helps in remembering information [from the audio input] better because it goes in to your long-term memory."
"it is a good idea for us the international students to get all the data [information from the audio input] that the person its asking for and also to make things clear."
"was a nice task. because improve my listening and i had to put my attention in that, and the results in the second part was better, more got it more details."
"i like the repetition because if you are confuse in one world and you listen twice you can understand what world your was confuse."

Note: The statements were not edited for grammar, spelling or punctuation

There were 7 (13.7%) positive mentions of writing out of the 51 total positive responses (4 from the Feedback group, 3 from the No-Feedback group). Table 4-55 provides examples of such positive opinions that reflect the writing skill.

Table 4-55. *Examples of positive mention of writing*

Participant's comment

“it helps to improve my written and get more ideas”
“It helps me to know what is my writing skills.”
“well is interesting because help me to writing to much better”
“The tasks helped me improve my writing. for the second time I wrote much better than the first time.”
“Is good for improve my English and be better in future in my essays.”

Note: The statements were not edited for grammar, spelling or punctuation

There were 14 (27.5%) positive mentions of both listening and writing out of the 51 total positive responses (10 from the Feedback group, 4 from the No-Feedback group). Table 4-56 provides examples of such positive opinions that reflect both listening and writing.

Table 4-56. *Examples of positive mention of both listening and writing*

Participant’s comment
“is very important because it help to improve listening and writing.”
“It is actually a great way to enhance my comprehension and writing skills.”
“this is good improves my English writing, listening. I like to do this task repetition in future classes.”
“in my opinion about task repetition it is helps me about my writing improvement and my listing part. I love this this task.”
“I think that is a good idea to perfect your listening and writing. That helps to understand more an idea.”

Note: The statements were not edited for grammar, spelling or punctuation

There were a total of 6 (11.8%) positive mentions of language skills more generally out of the 51 total positive responses (3 from the Feedback group, 3 from the No-Feedback group). Table 4-57 provides such examples of student responses.

Table 4-57. *Examples of positive opinions of task repetition (language skill benefits more generally)*

Participant’s comment
“in my opinion is good because the task help to know what you skills.”
“It was good because I could see how is my level of English”
“In my personal opinion it is very interesting because to improve my English. Thank YOU:)”
“In my opinion, it is a very good idea for international students to improve their English.”
“As a ESL student, I'm doing whatever I can to improve my language skills. I look forward to do the next one.”

Note: The statements were not edited for grammar, spelling or punctuation

Not all positive opinions included explanations, however. Out of the 51 total positive responses, there were a total of 12 (23.5%) positive opinions that did not include explanations (2 from the Feedback group, 10 from the No-Feedback group). Table 4-58 provides examples of such student responses.

Table 4-58. *Examples of positive opinions of task repetition (without explanations)*

Participant's comment
"if was good. everything was helpful"
"It was a good experience for me. I enjoyed it. Overall it was good."
"It is a good idea."
"Everything is fine for me"
"the test was good"

Note: The statements were not edited for grammar, spelling or punctuation

Within the positive opinions that did include explanations, there are several that included opinions about task repetition that are worth mentioning further. One example is "it was helpful and a new experience for me." In this case, this student emphasized the positivity in the novelty of this learning tool. Several opinions showed positive emotions in the responses. An example, besides the one provided in Table 4-57 with the smiling emoji, is "GOOD, I think they can inprove my writing and listening." This form of exclamation with capital letters suggests another example of positivity in a student's view of task repetition. Furthermore, the following two examples of opinions show how students felt task repetition helps enrich their vocabulary development: "they are helpful to understand any words that i missed" and "that was good because the task help me to know words meanings". These examples of opinions show a few of the ways that students perceive task repetition. At the same time, not all positive opinions that the students shared reflected specifically on their views of task repetition. Next are some examples of opinions that reflect other angles.

Within the positive opinions, there were replies that indicated overall a favorable view on the task while not commenting specifically on the repetition, but suggesting that they may need to do even more tasks to develop further. An example is "It is a good task for me, but I need more practice writing." In this case, this student identified a main area of skill improvement through more practice of task completion. Building on skill improvement, another student's response focused on the way that integrated skills help improve language proficiency: "I strongly agree with the method of listening in order to write an essay because in

my opinion I'll use two senses, and these senses working together can complement each other". Because this questionnaire was completed immediately after the final task repetition, it would seem that this student would likely find that combining multiple skills while also repeating the task would improve language development. Another example of a self-reflection statement that did not address the student's opinion about task repetition is "This is a good exercise for ESL student to do, but I don't have enough time to express my opinions about 3 questions [from the listening-to-write task]". It is clear that this student recognised that the needed skill for improvement is to be able to plan and complete a task within the allotted time (while indicating that they struggled with that). This student, like all participants in this study, had already completed a third performance of this task before this questionnaire was presented to them. However, it cannot be automatically assumed that this student would have found that additional repetitions would help improve language development.

Regarding the *negative opinions*, there were a total of 5 (8.5%) negative comments out of the 59 total responses received (2 from the Feedback group, 3 from the No-Feedback group). Table 4-59 provides these five negative opinions about task repetition.

Table 4-59. N*egative opinions of task repetition*

Participant's comment
"is too boring they could find another thing to listen."
"Writing a task was to boring"
"I dont think I like the repetition."
"not necessary"
"My overall opinion about task repetition is that I think that we should do something else because we already did that and this make the work boring"

Note: The statements were not edited for grammar, spelling or punctuation

Regarding the *mixed opinions*, there were a total of 3 responses, (5.1%) out of the 59 total responses received (3 from the Feedback group, 0 from the No-Feedback group). Table 4-60 shows the three mixed opinions to about task repetition.

Table 4-60. *Mixed opinions of task repetition*

Participant's comment
"A little bit boring but it is helpful"
"I feel sometimes is really helpful but sometimes is not"

"sometimes it is beneficial to repeat a task if it contains a lot of information that the brain can't process immediately, sometimes it is just a waste of time if the info is simple and easy to get at the first time."

Note: The statements were not edited for grammar, spelling or punctuation

4.3.3 Conclusion on student perceptions of task repetition

In conclusion to student perceptions of task repetition, students overall held positive views on the task used in the study and very positive views on listening-to-write tasks and task repetition more generally. Students particularly seemed to think that repeating a task improved their writing. Also, there was no statistically significant difference between the perception scores of students who had received feedback on their writing and those who had not. This was the case for their views on the task used in this study, on listening-to-write tasks more generally, as well as on task repetition more generally. Overall, students' levels of agreement on the statements (as positively worded) on the Likert scale were mainly "agree" and "strongly agree", with some tick boxes under "neutral." Similarly, overall, the majority of the students' opinions from the open-ended question about task repetition were positive regardless of the feedback condition. The students who provided explanations with their opinions about task repetition collectively provided a variety of reasons why they held positive views.

4.4 Chapter summary

In this chapter, I presented the results of my analyses of the effects of task repetition and feedback on CAF and knowledge summary and transfer displayed in students' writing produced for a listening-to-write task. For each measure, I presented a table to report the descriptive statistics based on time periods split out by feedback conditions. I then presented the comparative statistics for all of the measures through mixed-between-within analyses of variance (ANOVA) + interaction to determine if there is an effect of repetition and feedback condition. I also presented the results of students' perceptions about this task and task repetition based on their responses to Likert scale level of agreement to statements on post-task questionnaire. I showcased student perceptions about this task, as well as the extent to which they agree that writing listening-to-write tasks and repeating tasks improve their writing. Comparative statistics were used to determine the effect of feedback on student perceptions of task repetition. I provided tables and figures throughout this chapter to help display the results.

As per RQ1, there were interaction effects of repetition and feedback for several of the CAF measures. However, there were no interaction effects of repetition and feedback for knowledge summary and transfer.

As per RQ2, it was found that repeating a listening-to-write task has a positive effect on CAF in many respects and on knowledge summary and transfer. Namely, repetition had a positive effect on most of the CAF measures of students' written performances.

As per RQ3, there was a lack of effect on feedback on CAF and on knowledge summary and transfer.

As per RQ4a, students in both feedback-condition groups showed positive views on the task used in this study as well as on the integration of listening-to-write more generally (in terms of the extent to which these might help develop students' academic writing proficiency). As per RQ4b, there was no statistically significant difference in task repetition perception scores between the two feedback-condition groups. Additionally, the number of positive opinions that students shared regarding their views on task repetition in their responses to the open-ended question were very high and quite similar between the two feedback-condition groups.

5 Discussion

In this chapter, I present a summary of the results per research question, followed by a discussion of the findings in terms of the interaction effect of repetition and feedback (RQ1) in 5.1, effect of repetition (RQ2) in 5.2, effect of feedback (RQ3) in 5.3, and in terms of student perceptions about task repetition (RQ4a and RQ4b) in 5.4. I discuss the means for specific measures that were directionally consistent with the hypotheses. Similarly, I provide possible explanations where the findings trended opposite to the hypothesised directions. Further, I align the findings of my study with previous studies.

5.1 Interaction effect of repetition and feedback (RQ1)

RQ1 asked "Is there a change in listening-to-write task performance (CAF, knowledge summary and transfer) as a function of task repetition (Time) and feedback, such that the effect of task repetition (time) on the task performance depends on whether students receive feedback or not?"

The analyses in Chapter 4 revealed that there were interaction effects between repetition and feedback for CAF measures. However, there was no interaction effect between repetition and feedback for knowledge summary or knowledge transfer.

Next is a summary of the results in terms of interaction effects between repetition and feedback for CAF measures. As shown in Table 4-36, there were interaction effects between repetition and feedback for *seven* of the 16 CAF measures. At the same time, this table also displays the measures that did not show significant interaction effects, which suggests that the interaction between repetition and feedback was not even across the CAF dimensions. For the CAF measures where there was an interaction, all except preposition errors per 100 words had large effects; preposition errors per 100 words had a moderate effect.

Complexity

- average sentence length
- ratio of simple to complex sentences
- sophisticated words
- lexical sophistication

Accuracy

- preposition errors per 100 words

Fluency

- W/T

- W/EFT

In terms of hypothesised directions, improvements were anticipated across all CAF measures for both groups. However, data on this table showed that some of the outcomes only partially aligned with the predictions, i.e., there were significant declines or non-significant changes.

As shown in Table 4-37, simple effects of time (for the seven measures with interaction effects) were split out for the two student groups (feedback/no feedback). This table showed where the significant effects for the seven above-mentioned CAF measures were, as well as the size of the effects, along with descriptions of the directional trends that occurred for each group. Overall, the feedback group generally appeared to follow more consistent patterns of developmental trajectories; the no-feedback group showed more variation in patterns, including some stability as well as a few declines. As the directional trends suggest, the results were not always linear for all seven measures. In addition, Table 4-38 summarized significant improvements and declines, as well as non-significant differences across performances as split out by time and feedback condition for all CAF measures. The patterns presented in this table suggest that the feedback group achieved not only more consistent gains but maintained them across performances, whereas for the no-feedback group, the findings suggested more fluctuation. The data presented on Tables Table 4-37 and Table 4-38 show that feedback contributed toward steadier developmental trends compared to variation in trends where there was no feedback. Out of the seven measures, the following four are CAF measures where there were large effects for *both groups*:

Complexity

- average sentence length
- sophisticated words
- lexical sophistication

Accuracy

- preposition errors per 100 words

The two CAF measures where there was a large effect for the *feedback group* but no effect for the no-feedback group were: ratio of simple to complex sentences and W/EFT. The one CAF measure where there was a large effect for the *feedback group* and a moderate effect for the *no-feedback group* was W/T.

It is worth comparing the CAF measures where there was an interaction effect between repetition and feedback condition **and** a main effect of repetition or feedback, as opposed to

the CAF measures where there was an interaction effect but with **no** main effect. Out of the seven CAF measures that had an interaction effect, as shown in Table 4-39In terms of repetition (RQ2), the results in terms of CAF measures reported in section 4.1 show that students' EAP writing elicited by the listening-to-write task employed in this study statistically significantly improved with repetitions, as based on several measures. Table 4-39 displays a detailed overview of the effects of repetition on CAF performance indicators for which this was the case, i.e., a significant main effect of repetition, as well as for those for which it was not the case. As shown in Table 4-39, the directional arrows reflect the direction of numerical values. For errors per 100 words, errors per T-unit and the ratio of simple to complex sentences, decreases (↓) represent improvements; for the other measures, increases (↑) represent improvements; and "x" represents no significant change. This clarification ensures consistency with the patterns shown in Table 4-38. While both groups made improvements over time on various variables, a greater number of improvements were observed in the feedback group. Findings for the no-feedback group suggest more mixed outcomes, such as some significant declines or non-significant changes.

All hypothesised directions were set as predicted improvements shown by increases (⇑) or decreases (⇓) of values for the measures. In terms of results that aligned with the hypothesised improvement predictions, this table shows (✓) only when a statistically significant main effect confirmed the predicted improvement. For non-significant main effects, no alignment indicator is shown, and (✓) therefore indicates improvement supported by the overall main effect, not by isolated time interval comparisons, e.g., T1–T2, etc. The nine CAF measures that had a main effect of repetition are listed in blue., the following three are measures that also had an effect of repetition:

Complexity

- sophisticated words

Accuracy

- preposition errors per 100 words

Fluency

- W/EFT

Out of the seven CAF measures that had an interaction effect, the following four had no effect of repetition or feedback in isolation:

Complexity

- Average sentence length

- Ratio: simple to complex sentences
- Lexical sophistication

Fluency

- W/T

Several of the results that did not reach statistical significance did show small or moderate effect sizes. Such a pattern exemplifies the distinction between p-values and effect size. P-values are influenced by the sample size and variability; effect sizes signify the magnitude of the effect. In this study, I reported these findings for transparency, while at the same time, they should be interpreted with caution. Such findings might indicate possible trends. However, without replication of this study, these possible trends should not be interpreted as robust evidence, in line with recommendations for effect size interpretation in L2 research (Plonsky & Oswald, 2014).

In terms of the four above-mentioned CAF measures where there were interaction effects but **no** main effects, this seems to suggest that in order for students to write more complex sentences and more words, merely repeating the task or receiving feedback (in isolation without the other) is not always enough. Potentially, it suggests that students need to receive feedback after they repeat their writing performances, i.e., neither of the two interventions (repetition and feedback) worked alone by themselves for there to be statistical effects of these four measures; they needed to have both effects working together for there to be a pedagogic effect. For instance, feedback might have been effective when paired with enough frequent repetition, or vice versa, to work together such that the combined effect was more substantial than an effect of feedback condition or repetition in isolation. In this case, the relationship between feedback condition and repetition would be conditional with one impacting the effectiveness of the other.

Continuing with the four CAF measures (*average sentence length, ratio: simple/complex sentences, lexical sophistication, and W/T*) that needed to have both conditions working together for there to be an effect, I next discuss possible reasons why the whole group did not have main effects in isolation. It is important to note that, more generally, there are some commonalities between the following three measures: *average sentence length, ratio: simple/complex sentences,* and *W/T*. The commonality is that they relate to the ratio of words per chunk as well as the type of sentences, i.e., simple versus complex. The findings from my study suggest that there are some similarities between the results that occurred for the two groups for these measures. For example, a commonality between the results of the two

syntactic complexity measures, i.e., average sentence length and ratio: simple/complex sentences, is the feedback group started at a lower level of complexity at Time1 than did the no-feedback group. In fact, at Time1, the difference in the means for average sentence length and ratio of simple to complex sentences between the two groups was statistically significant (see Table 4-40). For W/T, the means for the feedback group at Time1 was somewhat lower, although not significantly lower than for the no-feedback group. Therefore, it seems that for these measures, overall, there was more scope for the feedback group to improve at the repeated performances than for the no-feedback group. In this case, for the no-feedback group, there was not enough maneuvering space for improvements in complexity despite some inexplicable drops at Time 2.

Next, I discuss the *four* CAF measures individually in terms of potential reasons why there were interaction effects without main effects of repetition or feedback.

Average sentence length. For average sentence length, for both groups together, there was no statistically significant effect of repetition. While overall the groups ended up at a similar level, the feedback group started off much lower than the no-feedback group, and then the feedback group wrote longer sentences at each repetition. Yet, on its own, the repetition effects did not lead to a statistically significant difference; the interaction effects led to significant differences and not the individual effects.

A possible element of the writing performances that may have led to longer sentences is the students were focusing on writing more explicitly, i.e., expressing ideas or information in detail in a clear way. For example, when they started to write longer sentences, they also had higher scores for knowledge transfer, i.e., more explicit in their writing, such as explaining things that college students should do to succeed and what they look for in an excellent professor. In trying to make the knowledge transfer score better, the consequence of that might have been that they ended up writing longer sentences. This suggests that the more that the students applied the information from the input and wrote more critically and explicitly than the previous times (i.e., throughout the repetitions, as both groups significantly improved in knowledge transfer at each time), the longer the sentences would be than if they were asked to merely summarize or paraphrase the input.

To determine the extent to which this correlation would be plausible, I computed a Spearman rank order correlation to assess the relationship between average sentence length and knowledge transfer scores at Times 1, 2, and 3. Findings from this test show that there is some correlation between the two measures. At Time1, there was a non-significant, weak, positive correlation between average sentence length and knowledge transfer score, $r(62) = [.23]$,

p=[.07]. At Time2, there was a non-significant, weak, positive correlation between average sentence length and knowledge transfer score, $r(62) = [.22]$, *p*=[.08]. However, at Time3, there was a significant, medium, positive correlation between average sentence length and knowledge transfer score, $r(62) = [.38]$, *p*=[.002]. These results suggest that for both groups taken together, students who practice more explicit writing also write longer sentences.

With the repetition, both groups knew the content from the listening input better than in the earlier performances. So they had more content and ideas to write about, thus this suggests that they could focus more on the writing and be able to write longer sentences as they had more content that they could cover as well as details they could add.

In terms of the feedback group, because they had feedback on grammatical structures, there were able to focus more on the content than on grammar. However, this group did not start to show a significant improvement in average sentence length until Time1-3, which may correlate with the significant, medium positive correlation at Time3.

In terms of the no-feedback group, average sentence length did not get better from baseline. In fact, in their second performance, they had declines, and then from Time2-3, they made significant improvements, but that improvement at Time3 was merely a return to where they started at Time1. This may suggest why the correlations between average sentence length and knowledge transfer were not significant at the start for the group.

Ratio: simple to complex sentences. For ratio of simple to complex sentences, the feedback group had written more simple sentences than complex sentences at Time1, then they wrote more complex sentences at each repetition. At the same time, the no-feedback group had an inexplicable complexity drop at Time2, i.e., they wrote more simple sentences than complex sentences at Time2, then at Time3, the ratio was the same as that of Time1. Meanwhile, the feedback group made steady decreases of simple sentences in proportion to complex (an improvement at each time). Because this measure is somewhat related to average sentence length, it would make sense that as the feedback group wrote more complex sentences than simple sentences that their sentences would be longer. Further, when the no-feedback group had an inexplicable increase in simple sentences at Time2, it would make sense that this group wrote shorter sentences at Time2. This measure needed the repetition to occur simultaneously with the feedback for there to be a pedagogic effect.

W/T. In terms of W/T, together, there was an increase from Time1 to Time3, but a decrease from Time1 to Time2. By feedback condition, the feedback group started at a somewhat lower means at Time1 than did the no-feedback group, then increased at each time. The no-feedback group, however, made a steady decrease from Time1 to Time2, with only a

slight increase from Time2 to Time3. Because the feedback group started out lower at Time1, there was more space for improvements than for the no-feedback group. This pattern of observation in this finding is similar to the findings for average sentence length and ratio of simple to complex sentences. With such elements in similarity, notably between average sentence length and W/T, it would be logical that the feedback group is drawing their attention to improving their writing in a systematic manner, i.e., increasing their amount of writing as they write longer sentences and fewer simple sentences in proportion to complex sentences. It appears that more repetitions may have resulted in a pedagogic effect.

Lexical sophistication. In terms of lexical sophistication, taken together and unlike with average sentence length, ratio: simple/complex sentences, and W/T, both groups had very similar means at each time. However, when split out by feedback condition, the feedback group had a slight decrease from Time1 to Time2, then a steady increase at Time3, thus a need for there to be an additional repetition and feedback (or multiple repetitions) for the feedback group to focus on lexical sophistication. The no-feedback group, however, had a steady increase from Time1 to Time2, then a steady decrease at Time3, yet still small differences overall for this measure.

Potentially, one can deduce that these four CAF measures were able to reach an interaction effect between repetition and feedback through the above-mentioned interdependence. However, in isolation, they did not have significant effects. Also, the majority of the CAF measures with interaction effects (regardless of whether there were main effects of repetition) were complexity ones. It cannot be fully excluded that sentence structure and vocabulary knowledge might have also somewhat improved due to the fact that all students (both feedback group and no-feedback group) were in an EAP course, which could presume better study habits than in a non-EAP course. In addition, they were participating in my study, which might have influenced the amount of time the students studied. However, we would not expect, nor would we normally see large changes in the course of one to two weeks.

5.2 Effect of repetition (RQ2)

RQ2 asked "Is there a change in listening-to-write task performance (CAF, knowledge summary, and transfer) as a function of task repetition (Time)? If so, in which direction?"

The analyses reported in Chapter 4 revealed that there were effects of repetition for the two groups for many CAF measures, as well as for knowledge summary and knowledge transfer. In terms of CAF measures, nine out of 16 went in the hypothesised direction, i.e. improvements with repetition. Both groups made significant improvements over time, though a larger number of improvements were observed in the feedback group. In terms of knowledge

summary and knowledge transfer measures, both went in the hypothesised direction, and both groups made significant improvements over time. These findings suggest that when students repeat a written performance, they improve in various facets of their writing, such as grammar, syntax, and content itself. Next are summaries and analyses of the results on the effects of repetition for CAF measures in 5.2.1 and knowledge summary and transfer in 5.2.2.

5.2.1 CAF and repetition

For **CAF measures**, the analyses described in Chapter 4 (Table 4-39) revealed that there were statistically significant improvements for both groups when repeating the task, with large effect sizes in more than half of the CAF measures (nine out of 16). The majority of these significant gains, i.e., six out of nine, concerned accuracy-related measures. The effects of repetition most generally aligned with the hypothesised directions, although this was not universal across all measures. Importantly, several measures did not show significant changes, therefore suggesting that repetition did not uniformly improve performance across all CAF dimensions. The non-significant findings showed that while repetition may play an important role in supporting accuracy, its effects on other aspects of CAF, such as several measures of complexity and fluency, appeared more limited or inconsistent. Taken together, the results should be interpreted cautiously, as the benefits of repetition, although evident, were not evenly distributed across the full set of CAF measures.

The majority, i.e., six, of those nine measures concerned accuracy:

Complexity

- sophisticated words
- lexical diversity

Accuracy

- Errors per 100 words
- E/T
- EFT/T
- Subject-verb agreement errors per 100 words
- Verb form errors per 100 words
- Preposition errors per 100 words

Fluency

- W/EFT

Also, when comparing the results between the global accuracy measures (errors per 100 words, E/T, and EFT/T) versus error type accuracy measures, all three global measures and three of the five error type measures (subject-verb agreement, verb form, and preposition errors per 100 words) had effects going in the hypothesised direction. In terms of complexity, there were significant effects for sophisticated words and lexical diversity; in terms of fluency, there was a significant effect for W/EFT. It is apparent that students improved most strongly in the accuracy dimension of their writing as they repeated their writing performances. It is further worth noting that for both groups, it is the accuracy dimension of CAF, i.e., grammatical structures, compared to the other CAF components, where there were the most significant improvements earlier on in the writing performances. A potential reason for this is students likely were focusing more of their energy on improving their grammar than on other facets of their writing during their repeated task performances. This finding regarding accuracy suggests that using separate sets of accuracy measures helps researchers identify the specific grammatical features where students improve most. Also, the opposite might be relevant for teachers in that it would be informative regarding where students need more work to improve. In this case, using only broad measures such as global accuracy measures does not capture this detail. My use of these two types of accuracy measures, i.e., global and by error type, aligns with Ortega (1999) who suggested that using both will give a more focused overview of the accuracy findings. Further, by using error-type measures, they help researchers make direct comparisons between studies that use similar measures.

In addition to the above-mentioned accuracy findings, the mean number of sophisticated words and lexical diversity had significant effects with repetition., i.e., improving with repetition. This occurred even when the task did not explicitly instruct students to use advanced and varied vocabulary. Also, the task did not explicitly instruct any particular aspects that needed paying attention to, apart from not answering each question separately, which suggested the need to focus on creating coherence in covering the three questions in one essay. A possible reason for this finding is the repeated listening input helped enable them to retain more words as they became familiar with the task, which allowed them the opportunity to practice using more of the vocabulary they would have recently retained. Additionally, it might be that the repetitions meant they were more familiar with the content and might have had more cognitive space to paraphrase and express the ideas with richer language.

In terms of the fluency measure, repetition had an effect on W/EFT in the writing performances. A possible reason why this measure improved with repetition as opposed to W/T, which did not significantly improve with repetition, is that there is an accuracy

component to calculate words per error-free T-unit, and we have seen above that repetition positively impacted on several accuracy features. The W/T measure, on the other hand, as defined earlier, measures the number of words per T-unit but does not have an accuracy inclusion. This finding suggests that as students improved their grammar, they were able to write more sentences such that there were more T-units that did not contain the grammatical errors under scrutiny in this study.

Regarding W/T, the task did not specify the text length, only the time limit, which is a likely reason that W/T did not have significant effects of repetition for both groups taken together even though there were significant improvements for the feedback group from Times 2-3 and 1-3, whereas for the no-feedback group, there were significant declines from Times 1-2 and 1-3. Had this task required a word count range, there possibly would have been a significant effect on the number of words or word ratios for both groups taken together. Had that been the case, the students might have started writing in a more condensed manner to stay within the word limit.

Despite that the descriptive statistics show that the majority of the CAF measures improved at each performance, some very steadily, there are several possible reasons why not all CAF measures had significant effects of repetition. One possible reason is that the observed difference may have happened by chance. Another possible reason is there was too small a change between each time to reflect a significant difference, for example, the mean verb tense errors per 100 words at Time1 and Time2 were quite similar even though seemingly improving just slightly based on the raw numbers. There is an additional likely explanation. According to various SLA researchers, there are some grammatical structures (such as articles and verb tenses) that are notoriously challenging for many L2 learners, particularly if their L1 has a different grammar system (e.g. Ionin & Wexler, 2002; Ionin & Wexler, 2003; Dominique et al., 2017; Murakami & Alexopoulou, 2016).

It is important to note where there were statistically significant improvements between written performances. It is equally important to note where there were significant improvements at each time for both groups as opposed to just one group but not the other. As shown in Table 4-39, there were effects of repetition, and the following shows when the significant improvements occurred:

Significant improvements for both groups at each time. Two measures where there were significant improvements at each time for both groups are *errors per 100 words* and *E/T*. This suggests that, overall, students make fewer grammatical errors as they gain more writing

practice, whether they receive feedback or not. However, looking into the specific types of errors, the picture differs depending on the kind of error.

Significant improvements for both groups at only T2-T3 and T1-T3. there were four measures where there were significant improvements for both groups at only T2-T3 and T1-T3, but not yet at T1-T2: *lexical diversity, EFT/T, subject-verb agreement,* and *verb form errors*), This further suggests that students improve in certain aspects of their writing as they gain more writing practice whether they receive feedback or not, while at the same time, it may take more than one repetition, i.e., more than two writing performances for some of the improvements to occur.

Significant improvements with differing patterns. In contrast, for three measures, the improvement patterns with repetition differed between performances. Two error measures that do not share the similarities across the groups are *preposition errors per 100 words* and *W/EFT*. For *preposition errors per 100 words*, for the *feedback group*, there were significant improvements at each time; for the *no-feedback group*, the significant improvements occurred at only T2-T3 and T1-T3. For *W/EFT*, for the *feedback group*, there were significant improvements at each time; for the *no-feedback group*, there were no significant improvements with repetition. This indicates that the group that received feedback was able to show improvements in their writing earlier on than did the no-feedback group. Another mixed picture was found for one complexity measure, *sophisticated words*. Namely, for the *feedback group*, there were significant improvements at each time; for the *no-feedback group*, the significant improvements occurred at only T1-T2 and T1-T3.

As shown in Table 4-39, the following shows various patterns in terms of when significant improvements and non-significant differences occurred:

Non-significant differences for both groups from T1-T2, then significant improvements from T2-T3 and T1-T3. There were four measures that showed non-significant differences from T1-T2, then significant improvements from T2-T3 and from T1 to T3: *lexical diversity, EFT/T, subject-verb agreement errors,* and *verb form errors.*

Non-significant differences for both groups at each time. Three measures that showed no significant differences for either group at any of the times were *C/T, verb tense errors,* and *article errors*.

Next, I draw from Table 4-38 to identify where the information suggests there were trade-off effects, and lack thereof.

From Time1 to Time2, for both groups, as shown in Table 5-1 below, which shows the relevant data from the broader Table 4-38 that suggest partial trade-offs between accuracy

being the strongest CAF dimension versus some complexity measures. It is very important to note that in terms of accuracy, we are looking at main measures of accuracy, i.e. the global measures: errors per 100 words, E/T, and EFT/T.

For the *feedback group*, the data suggest that the partial trade-off effect occurred between accuracy and several specific complexity measures (average sentence length, C/T, lexical diversity, and lexical sophistication). Because only two of the six complexity measures significantly improved, i.e., ratio of simple to complex sentences and sophisticated words, this makes this a partial trade-off, thus, this therefore suggests it partially backs the trade-off hypothesis at the second performance in terms of competition between accuracy and complexity (for specific measures).

For the *no-feedback group*, findings show that a stronger trade-off effect occurred between accuracy (errors per 100 words and E/T) versus complexity and fluency (where there was a significant decline in one complexity measure, average sentence length, and one fluency measure, W/T), suggesting that even if students are given two opportunities to write the same task with or without feedback, then some improvements will begin to occur at the expense of other aspects of their writing. What makes the trade-off even more visible in the no-feedback group than in the feedback group is there are significant declines that accompanied the significant improvements. For the feedback group, the significant accuracy improvements occurred at the expense of non-gains in some of the other CAF measures. In this case, the significant accuracy improvements were not accompanied by significant declines.

Table 5-1. *Written performances: Improvements/Declines/Similarities (for Time1-2 by feedback condition)*

	Feedback Group		No-Feedback Group
Measure	Time 1-2		Time 1-2
Average sentence length	*Non-sig.*		Declined
Ratio: Simple to complex sentences	Improved		*Non-sig.*
Clauses per T-unit (C/T)	*Non-sig.*		*Non-sig.*
Lexical diversity	*Non-sig.*		*Non-sig.*
Sophisticated words	Improved		Improved
Lexical sophistication	*Non-sig.*		*Non-sig.*
Errors/100 words	Improved		Improved
Errors per T-unit (E/T)	Improved		Improved
Error-free T-units/T-unit (EFT/T)	*Non-sig.*		*Non-sig.*
Subject-verb errors/100 words	*Non-sig.*		*Non-sig.*
Verb tense errors/100 words (VT)	*Non-sig.*		*Non-sig.*
Verb form errors/100 words (VF)	*Non-sig.*		*Non-sig.*
Preposition errors/100 words	Improved		*Non-sig.*
Article errors/100 words	*Non-sig.*		*Non-sig.*
Words per T-unit (W/T)	*Non-sig.*		Declined

Words per error-free T-unit (WEFT)	Improved		*Non-sig.*

Notes:

1. Decreases in errors/100 words, E/T, or in ratio of simple to complex sentences are improvements; increases are declines.

Color codes:

2. Blue = scores have significantly improved between the specified times.
3. Red = scores have significantly declined
4. Black = no significant differences in the scores

From Time2 to Time3, as shown in Table 5-2 below, which shows the relevant data from the broader Table 4-38 for both groups, overall, there were more significant improvements in accuracy compared to those from Time1 to Time2. For the *feedback group*, the data suggest that a partial trade-off effect occurred between the significant accuracy, fluency, and several complexity improvements versus specific complexity measures (average sentence length, ratio of simple to complex sentences, and C/T). Several more of the accuracy measures improved beyond errors per 100 words and E/T, i.e., EFT/T, subject-verb agreement, verb form, and prepositions, as did two new complexity measures beyond sophisticated words, i.e., lexical diversity and lexical sophistication.

For the *no-feedback group*, findings show that a stronger trade-off effect occurred for this group than for the feedback group. This potential trade-off was between accuracy and some complexity measures (average sentence length and lexical diversity) where there were significant improvements, versus lexical sophistication, where there was a significant decline. Like from Time1 to Time2, accuracy is stronger than complexity. This suggests an earlier point that as students gain more writing practice, they make further improvements, most notably in accuracy and in some complexity measures than if they had only one opportunity to repeat the task. It is important to note, as mentioned earlier, that for the no-feedback group, the significant improvement in average sentence length was not an improvement but rather a significant improvement after a significant decline, i.e., it was merely a return to baseline

Table 5-2. *Written performances: Improvements/Declines/Similarities (for Time2-3 by feedback condition)*

	Feedback Group		No-Feedback Group
Measure	Time 2-3		Time 2-3
Average sentence length	*Non-sig.*		Improved
Ratio: Simple to complex sentences	*Non-sig.*		*Non-sig.*
Clauses per T-unit (C/T)	*Non-sig.*		*Non-sig.*
Lexical diversity	Improved		Improved
Sophisticated words	Improved		*Non-sig.*
Lexical sophistication	Improved		Declined

Errors/100 words	Improved		Improved
Errors per T-unit (E/T)	Improved		Improved
Error-free T-units/T-unit (EFT/T)	Improved		Improved
Subject-verb errors/100 words	Improved		Improved
Verb tense errors/100 words (VT)	*Non-sig.*		*Non-sig.*
Verb form errors/100 words (VF)	Improved		Improved
Preposition errors/100 words	Improved		Improved
Article errors/100 words	*Non-sig.*		*Non-sig.*
Words per T-unit (W/T)	Improved		*Non-sig.*
Words per error-free T-unit (WEFT)	Improved		*Non-sig.*

Notes:

1. Decreases in errors/100 words, E/T, or in ratio of simple to complex sentences are improvements; increases are declines.

Color codes:

2. Blue = scores have significantly improved between the specified times.
3. Red = scores have significantly declined
4. Black = no significant differences in the scores

From Time1 to Time3, as shown in Table 5-3 below, which shows the relevant data from the broader Table 4-38 for both groups overall, there continued to be significant improvements in the same accuracy measures as at Time2 to Time3.

For the *feedback group*, the data show significant improvements in the three CAF dimensions. Concurrently as the same accuracy and fluency measures as from Time2 to Time3 significantly improved, there also were further significant improvements in complexity, i.e., average sentence length and ratio of simple to complex sentences, thereby the CAF measures were no longer competing with one another, i.e. the trade-off effects disappeared at the third performance, thus suggests it backs Sample and Michel's (2014) finding that the CAF trade-off effects disappear by the third performance.

For the *no-feedback group*, findings show that a strong trade-off effect occurred between accuracy and specific complexity measures (lexical diversity and sophisticated words) versus the lexical sophistication complexity measure and the W/T fluency measure (where there were significant declines). These differences between the two groups may indicate that if students receive feedback as they repeat their writing performances, then they make improvements in more facets of their writing, hence the disappearance of the trade-off, which happened for the feedback group. Additionally, for the no-feedback group, because errors in writing decreased significantly (an improvement) while the complexity measure, lexical sophistication and the fluency measure, W/T, decreased significantly (a decline), this suggests that as students learn how to correct language, this may occur at the expense of focusing simultaneously on using more advanced vocabulary. At the same time, there was no longer a significant difference in average sentence length, thus this group was not writing longer sentences.

Table 5-3. *Written performances: Improvements/Declines/Similarities (for Time1-3 by feedback condition)*

	Feedback Group		No-Feedback Group
Measure	Time 1-3		Time 1-3
Average sentence length	Improved		*Non-sig.*
Ratio: Simple to complex sentences	Improved		*Non-sig.*
Clauses per T-unit (C/T)	*Non-sig.*		*Non-sig.*
Lexical diversity	Improved		Improved
Sophisticated words	Improved		Improved
Lexical sophistication	Improved		Declined
Errors/100 words	Improved		Improved
Errors per T-unit (E/T)	Improved		Improved
Error-free T-units/T-unit (EFT/T)	Improved		Improved
Subject-verb errors/100 words	Improved		Improved
Verb tense errors/100 words (VT)	*Non-sig.*		*Non-sig.*
Verb form errors/100 words (VF)	Improved		Improved
Preposition errors/100 words	Improved		Improved
Article errors/100 words	*Non-sig.*		*Non-sig.*
Words per T-unit (W/T)	Improved		Declined
Words per error-free T-unit (WEFT)	Improved		*Non-sig.*

Notes:

1. Decreases in errors/100 words, E/T, or in ratio of simple to complex sentences are improvements; increases are declines.

Color codes:

2. Blue = scores have significantly improved between the specified times.
3. Red = scores have significantly declined
4. Black = no significant differences in the scores

As discussed earlier, it is important to note that various studies have used different CAF measures from one another, some that exemplified trade-off effects and others not. To that end, this makes direct comparisons between studies difficult to make. Next are findings from my study that partially back the trade-off hypothesis at the second performance in terms of competition between complexity (for specific measures) and accuracy.

The trade-off effects mentioned above for the second performance in my study align with the findings of several empirical task repetition studies introduced in sections 2.3.4 and 2.3.5 in that there was, indeed, the presence of a trade-off. However, the direction of the trade-off in my study was opposite of the trade-offs in some of the previous studies. While complexity was stronger than accuracy in some previous studies, accuracy was the stronger dimension over complexity during the trade-off in my study, thus results from my study only partially aligned with these complexity-accuracy trade-off results. For instance, a written task repetition study where there was a trade-off is Jung's (2013) study (section 2.3.5) which found that for some of the groups, there was a trade-off between accuracy (errors per 100 words and error-free clauses) and complexity (C/T). In terms of the basis of the groupings, they were

placed in 4 groups: same task repetition with feedback; repetition without feedback; feedback without repetition; no repetition no feedback. In terms of trade-off's, the findings did not suggest trade-off's for the repetition groups, i.e., the groups that repeated the same tasks. However, for the no-repetition groups, i.e., the groups that did not repeat the same task, there were trade-off's in accuracy and fluency at the expense of complexity.

Similar trade-off's at Time 2 in my study have also been observed in the oral task repetition literature. For instance, in Bygate's (2001) study, there was a trade-off effect between accuracy (E/T) and complexity (pauses per T-unit and W/T), in which complexity and fluency were stronger than accuracy in the trade-off. Although W/T is a fluency measure, Bygate stated that "To some extent [W/T] might be thought of as a covert fluency measure… number of W/T reflects more than just speed: it also involved the extent to which lexical accessing can be managed according to basic syntactic parameters – cognitive capacity" (p. 34). The students improved in the complexity aspect of their speech production, but they made more grammatical errors while they were doing so. In Ahmadian and Tavakoli's (2011) study, there was a trade-off between complexity (subordination, ratio of clauses to AS units, and variety of verb forms used) and accuracy (error-free clauses and correct verb form usage), where complexity was stronger than accuracy. The students in the repetition group improved in complexity and fluency (number of meaningful syllables per minute) but there was no significant change in accuracy, thus complexity and fluency being stronger than accuracy. In Muhammadpour et al.'s (2023) study, there was a trade-off between complexity (amount of subordination and lexical diversity) and accuracy (error-free clauses), where complexity was stronger than accuracy.

Next, I sum up what my study adds to the existing knowledge base of trade-off patterns with complexity and accuracy dimensions competing with one another at the second performance in an integrated listening-to-write task repetition. A difference between many studies compared to mine is that they only had one repetition and thus there was no third performance. So, I am not able to state whether the trade-off effects would have disappeared in subsequent performances in those studies. Because my study had two repetitions, i.e., three performances, in this case, my study provided results on the extent to which the trade-off effects disappeared at the third performance.

A study that did have two task repetitions is Sample and Michel (2014), and my study's findings partially align with theirs. Namely, for the feedback group in my study, there was the presence of a trade-off effect at Time2, then the students improved in the same areas at Time3. In Sample and Michel's (2014) oral repetition study, at Time2, students used more complex

structures (complexity), and they made more grammatical errors (accuracy) such as agreement errors, article errors, as well as a decrease in error-free clauses. In terms of complexity and fluency, students who used more elaborate words made more pauses, therefore complexity was stronger than both accuracy and fluency in this trade-off. At Time3, the trade-off effects disappeared. Similar to Sample and Michel's findings, the trade-off effects in my written task repetition study disappeared for the feedback group in my study, which suggests that this finding from Sample and Michel's oral task repetition study can extend to studies on writing.

Taking these points a step further, despite the fact that the no-feedback group had not received feedback in my study, this group nevertheless made significant improvements in accuracy (see Table 4-38). However, unlike the feedback group, the no-feedback group did not make significant improvements in complexity to the same larger extent that the feedback group did. Because the feedback group made significant improvements in more CAF dimensions than did the no-feedback group, this means that the feedback group was already aware of the accuracy measures that needed improvement. This latter point suggests the possibility that by already knowing the grammatical mistakes that the feedback group participants needed to correct because such mistakes had been pointed out in their feedback, this allowed the feedback group to devote some more time to improving the complexity and fluency dimensions. The no-feedback group instead needed to determine on their own the grammatical features of their writing that needed improvement, which might have reduced the amount of time that they could focus on all other CAF features. This partially aligns with my hypothesis that by receiving feedback, it will help students improve their writing. At the same time, because the no-feedback group made significant improvements in accuracy, this would suggest that even without feedback, repetition helps students improve in their writing performances with at least some trade-off effects or part thereof.

Having discussed the trade-off effects observed in the data based on the quantitative measures, I will now illustrate what this looks like in an actual sample of a student's three written performances where there were trade-off effects between accuracy and complexity. First, I will describe the patterns in general. Then, I show the figures, and finally, below the figures, I discuss the nature of the student's written performances in terms of the trade-off effect in more detail.

Figure 5.1 shows Student 10's texts at all three performances in terms of accuracy features at Times1, 2, and 3, with accuracy errors highlighted in yellow. Student 10 was in the feedback group. As introduced in the methodology chapter in section 3.5.3.1**,** the grammatical error type measures focused on in this study were subject-verb agreement, verb tense, verb

form, prepositions, and article errors per 100 words. The visual pattern that can be observed from Figure 5.1 is that there are fewer highlighted words with each repetition, which exemplifies accuracy improvements at each time, i.e. fewer errors per 100 words.

Figure 5.2 shows Student 10's texts at all three performances in terms of sophisticated words (highlighted in yellow), a measure used for lexical sophistication, a measure of the complexity component of CAF. Sophisticated words, as introduced in the methodology chapter in section 3.5.3.1, are identified by the Lextutor lexical analysis webtool. The visual pattern in Figure 5.2 shows fewer highlights in the second compared to the first performance, thus a decrease of sophisticated words (a decline in this complexity measure) at Time2, and then more highlighted words and thus a steady increase of sophisticated words at Time3. By viewing Figure 5.1 and Figure 5.2 in combination, the reader can see the trade-off effect between accuracy (increasing) and complexity (decreasing at the expense of accuracy increases) at the second performance.

Figure 5.1. *Student 10's Repeated Writing Samples to Exemplify Trade-Off Hypothesis (Accuracy)*

Time 1

Success in Your Class

In the audio the professor was talking about of ESL classes. His name was Karl. His classes are at 3:15 to 4:50 at the room 405. the name of this classes was Communication 311. This classes could be take 2 months. In the classes his students has to buy a book, but unfortunately the book is not coming yet. Soon you're able to find it in the bookstore. His office hours are from 1:00 P.M to 2:00 P.M only wednesdays, but you're able to make an appointment in any day. I think that an excellent professor have to be friendly, look professional and humane. If you can be humane you can assimilate more others people and also I think that an excellent professor have to be good person (and not obnoxious).

Time 2

Success in Your Class

Karl is a professor. He was talking about his classes. his classes are at 3:15 to 4:50 in room 405, classes Communication 311. That classes could be take 2 months. In the class his students has to buy a book, but unfortunately the book is not coming yet, but they're able to find it in the bookstore.

His office hours are from 1:00 P.M to 2:00 P.M only Wednesdays, but they are able to make an appointment anytime. I think an excellent professor has to be very friendly, look professional and be humane. If you can be friendly and positive people you can assimilate more on others people. Be a good professor mean that you have to be really great and you have to explain clear so easy to understand.

Time 3

Success in Your Class

Professor Karl was talking about classes. The professor said classes are at 3:15 to 4:50. Finally he told the students what they will be doing in the class. I think what should be good is I would like the professor tell about his self. I believe I need to trust a professor who meets with the class. His office hours are from 1:00 to 2:00 on Wednesdays but can make an appointment anytime. He shows students his transparent and trustworthiness and can express their self to him. I also want the students express themselves with others in the class. In my opinion it can be convenient for students to speak with other people, then their voices be heard and it will build their self-esteem. I think building self-esteem is a great confident builder and the students might think they have a loud voice and is needed in a workplace because you explain clearly, easy to understand.

Note: Not edited for grammar, spelling, or punctuation. Selected grammatical errors are highlighted in yellow.

Figure 5.2. Student 10's Repeated Writing Samples to Exemplify Trade-Off Hypothesis (Complexity)

Time 1

Success in Your Class In the audio the professor was talking about of ESL classes. His name was Karl. His classes are at 3:15 to 4:50 at the room 405. the name of this classes was Communication 311. This classes could be take 2 months. In the classes his students has to buy a book, but unfortunately the book is not coming yet. Soon you're able to find it in the bookstore. His office hours are from 1:00 P.M to 2:00 P.M only wednesdays, but you're able to make an appointment in any day. I think that an excellent professor have to be friendly, look professional and humane. If you can be humane you can assimilate more others people and also I think that an excellent professor have to be good person (and not obnoxious).
Time 2
Success in Your Class Karl is a professor. He was talking about his classes. his classes are at 3:15 to 4:50 in room 405, classes Communication 311. That classes could be take 2 months. In the class his students has to buy a book, but unfortunately the book is not coming yet, but they're able to find it in the bookstore. His office hours are from 1:00 P.M to 2:00 P.M only Wednesdays, but they are able to make an appointment anytime. I think an excellent professor has to be very friendly, look professional and be humane. If you can be friendly and positive people you can assimilate more on others people. Be a good professor mean that you have to be really great and you have to explain clear so easy to understand.
Time 3
Success in Your Class Professor Karl was talking about classes. The professor said classes are at 3:15 to 4:50. Finally he told the students what they will be doing in the class. I think what should be good is I would like the professor tell about his self. I believe I need to trust a professor who meets with the class. His office hours are from 1:00 to 2:00 on Wednesdays but can make an appointment anytime. He shows students his transparent and trustworthiness and can express their self to him. I also want the students express themselves with others in the class. In my opinion it can be convenient for students to speak with other people, then their voices be heard and it will build their self-esteem. I think building self-esteem is a great confident builder and the students might think they have a loud voice and is needed in a workplace because you explain clearly, easy to understand.

Note: Not edited for grammar, spelling, or punctuation. Sophisticated words are highlighted in yellow.

More specifically, as shown in Figure 5.1, the number of words in the written performance, 134, remained the same at Time1 and Time2, then increased to 160 at Time3. In most cases, the various types of accuracy errors decreased or disappeared with repetition (except for verb form errors which increased at each time). At Time1, Student 10 made 11 grammatical errors. The majority of the errors were prepositions and subject-verb agreement – five and three errors, respectively. Eleven errors out of 134 words equals 8.2 errors per 100 words. There were three further errors: one verb form and two article errors. The one verb form error was due to the use of the word "be" that was not needed in "could be take 2 months." In terms of the preposition errors, an example was the placement of a preposition where it was not needed: "talking about of ESL classes." Another example pertained to time frame: "at 3:15 to 4:50." It is interesting to note that this student had the correct preposition in "from 1:00 P.M. to 2:00 P.M." An additional example of a preposition error pertained to a collocation: a preposition was missing in "assimilate more others people." An example of an article error was one that was placed where it was not needed ("at the room 405").

According to some SLA researchers, (e.g. Kellerman, 1983; Ionin & Wexler, 2002), in terms of language transfer from a ESL learner's L1 and language visibility, students may visualise structures from their L1 that do not align with the L2, for example, resulting in errors

such as “at 3:15 to 4:50”, “talking about of ESL classes,” or the missing the preposition after “assimilate more.” These preposition errors exemplify potential L1 transfer, such that there are misuses and omissions of prepositions. Conversely, ESL learners whose L1 follows similar preposition rules as the L2 might make fewer preposition errors (Chodorow et al., 2007). In terms of article usage errors, in a similar vein, an ESL learner whose L1 does not have articles, or that has a different structure than articles in the L2, often use articles incorrectly, (e.g. Ionin & Wexler, 2003). For example, articles such as “the” is often overused when referring to a location or when used in a prepositional phrase (e.g. “at the room 405”).

At Time2, Student 10 made six grammatical errors, so five fewer than at Time1. The strongest improvement was the reduction of preposition errors from five down to two. The number of subject-verb agreement errors remained at two, and the number of verb form errors increased from one to two. The article errors disappeared at Time2. Six errors out of 134 words equals 4.5 errors per 100 words. In terms of the two preposition errors, there is one repeated error from Time1 (at 3:15 to 4:50”), and the other error is an incorrect use of preposition for the collocation “assimilate more on others people” that should be replaced by “with.” This student repeated the same subject-verb agreement mistake from Time1, “students has.” In terms of verb form, one of the errors was repeated from Time1 – “could be take” –, while the other error – “Be a good professor” – needs to change to “To be” or “Being.”

At Time3, Student 10 wrote more words, an increase from 134 to 160, yet made fewer grammatical errors than at Times1 and 2: a reduction from 11 to six to four errors at Time3. Four errors out of 160 words equals 2.5 errors per 100 words. Preposition errors improved at each time: a reduction from five (Time1) to two (Time2) to one error at Time3. The one prepositional error was a repetition from the previous texts, “at 3:15 to 4:50.” Verb form errors, however, increased at each time from 1 to 2 to 3 at Time3. Two examples of verb form errors pertained to missing the “to” to make the verbs infinitive, “I would like the professor tell about” and “the students express themselves.”

To sum up for the three written performances in terms of accuracy for Student 10, this student overall made steady improvements in errors per 100 words at each time. This student reduced the number of subject-verb agreement and preposition errors. There were no verb tense errors, and the number of article errors was minimal and disappeared at Time2 and stayed so at Time3. Interestingly, Student 10 made more verb form errors at each time, with various causes for each error. Some of the same errors were repeated across the performances.

Figure 5.2 (above) shows Student 10’s texts at all three performances in terms of sophisticated words.

The text at Time1 shows nine sophisticated words out of 134 words, of which six of those words were different from one another. The sophisticated words were audio, professor, communication, humane, assimilate, and obnoxious. “Professor” was used three times.

The text at Time2 shows a decrease in sophisticated words from nine to six out of 134 words, in which 4 of those words were different but none of the words were new to Time2. “Professor” was used three times. This decrease in sophisticated words occurred at the same time that this student made improvements in accuracy, i.e., fewer errors per 100 words. This suggests a trade-off effect between accuracy which improved from Time1 to Time2 and this specific measure of lexical complexity which declined from Tima1 to Time2 .

The text at Time3 shows a steady increase to 10 sophisticated words out of 160 words, in which six of those words were different and five of those words were new to Time3. “Professor” was used four times; “esteem” was used two times. The new words were transparent, trustworthiness, convenient, confident, and esteem. However, because some of the sophisticated words were repeated despite an increase in the number of sophisticated words, it would be questionable to suggest that trade-off effects between accuracy and complexity, specifically sophisticated words (lexical sophistication), disappeared.

It is worth noting that although there are overall findings from my study, there are also variations that deviate from my overall conclusions about trade-off in Student 10’s texts. Specific measures of complexity followed slightly different patterns for Student 10’s texts. For example, lexical diversity and average sentence length remained somewhat similar at each time. This variation exemplifies something about learner development, i.e., not linearly in all aspects at the same time.

Simultaneously, Student 10’s knowledge summary and transfer improved across the repetitions. The knowledge summary score increased from one to two to three; knowledge transfer score of 1 remained the same at Times1 and 2, then increased to three at Time3. This shows an example of a way that my research demonstrates that with multiple fine-grained measures, I can uncover a range of improvements that would not be uncovered if I used only CAF measures. This latter point aligns with Qin and Liu (2024) who suggest using communication/content/function-related (CCF) performance measures to complement CAF measures.

We also have to keep in mind that results from the statistical procedures show us tendencies of improvement through repetition across groups of students. However, within those groups, there are often individuals who deviate from the common trend by not benefitting from the repetition. Nevertheless, these are the minority. While Student 10’s

writing samples exemplify some CAF trade-off effects, not all students' writing performances showed improvements in CAF. Next is an example of Student 55, also from the feedback group, whose writing exemplifies an exception to those findings of significant improvement, i.e., who does not match the common trend of benefitting from the repetition by making improvements in their writing.

Figure 5.3 shows Student 55's texts at all three performances where there were similarities in terms of the accuracy dimension of CAF in writing performances at each time, i.e., the number of grammatical errors per 100 words (highlighted in yellow) under scrutiny for this study were similar at each time, i.e., the actual number of errors were proportionally similar at each time. Figure 5.4 shows Student 55's texts at all three performances where there were several similarities in terms of sophisticated words (highlighted in yellow), i.e., there were similar numbers of sophisticated words across the three writing samples. By viewing Figure 5.3 and Figure 5.4 together, the reader can see there was no trade-off effect.

Figure 5.3. *Student 55's Repeated Writing Samples to Exemplify Similarities in Performance (Accuracy)*

Time 1

Success in Your Class

Today, we listen to the professor tell the students the class schedule. He teaches at 3:11pm, 405 in the LPA. Tuesday and Thursday in the class, he need students online brought the books and bring to the class. He gave E-mail to the students and told the students could sent E-mail to him. The class have the examination in the middle and last two months of study. Look for an excellent qualities professor, I think that excellent professors to help my writing, reading and after essays went much high.

Time 2

Success in Your Class

Today, we are listen to the professor Carl talk about the class schedule to students. He name is Carl. He teaches in 3:11pm, on 405 in the LPA room. Tuesday and Thursday, he tell students online bought the books and bring to the class. He gives E-mail to students and told the students sent E-mails to him in 1PM to 2PM. The class has the examinations in the middle and end of two months of study. Look for an excellent quality in professor, I think a excellent professor help my writing, reading and essay get much better.

For me, this is good for my writing, reading, and listening. I believe means good jobs, better lifestyle, and education, you can become successful.

Time 3

Success in Your Class

Today, we listening to the professor Carl told about the class schedule. He teaches at 3:50-4:50 pm in 405 in the LPA. Tuesday and Thursday in the class, he needs students online to bought the books to bring to class. He E-mailed and tell the students to E-mail his office in 1 to 2PM in Wednesday. The class taking the examination in the middle and last two months of study. I thought excellent professors to helps my writing, reading and essay improve more.

Students have good writing and good reading. That means a professor needs to explaining. I think an excellent professor like a dictionary can help student's grammar to be successful.

I study a language class, study Monday – Friday. I will to continue to study here. For me, this is good for my writing, reading, and listening. I hope get better jobs, life, and education to be successful.

Note: Not edited for grammar, spelling, or punctuation. Selected grammatical errors are highlighted in yellow.

Figure 5.4. *Student 55's Repeated Writing Samples to Exemplify Similarities in Performance (Complexity)*

Time 1

Success in Your Class

Today, we listen to the professor tell the students the class schedule. He teaches at 3:11pm, 405 in the LPA. Tuesday and Thursday in the class, he need students online brought the books and bring to the class. He gave E-mail to the students and told the students could sent E-mail to him. The class have the examination in the middle and last two months of study. Look for an excellent qualities professor, I think that excellent professors to help my writing, reading and after essays went much high.

Time 2

Success in Your Class

Today, we are listen to the professor Carl talk about the class schedule to students. He name is Carl. He teaches in 3:11pm, on 405 in the LPA room. Tuesday and Thursday, he tell students online bought the books and bring to the class. He gives E-mail to students and told the students sent E-mails to him in 1PM to 2PM. The class has the examinations in the middle and end of two months of study. Look for an excellent quality in professor, I think a excellent professor help my writing, reading and essay get much better.

For me, this is good for my writing, reading, and listening. I believe means good jobs, better lifestyle, and education, you can become successful.

Time 3

Success in Your Class

Today, we listening to the professor Carl told about the class schedule. He teaches at 3:50-4:50 pm in 405 in the LPA. Tuesday and Thursday in the class, he needs students online to bought the books to bring to class. He E-mailed and tell the students to E-mail his office in 1 to 2PM in Wednesday. The class taking the examination in the middle and last two months of study. I thought excellent professors to helps my writing, reading and essay improve more.

Students have good writing and good reading. That means a professor needs to explaining. I think an excellent professor like a dictionary can help student's grammar to be successful.

I study a language class, study Monday – Friday. I will to continue to study here. For me, this is good for my writing, reading, and listening. I hope get better jobs, life, and education to be successful.

Note: Not edited for grammar, spelling, or punctuation. Sophisticated words are highlighted in yellow.

Figure 5.3 shows Student 55's writing samples for Times1, 2, and 3, and the samples exemplify similarities of the accuracy at each time, i.e., number of grammatical errors for the measures under scrutiny for this study proportionally the same. While some specific types of errors disappear when the task was repeated, for example, subject-verb agreement and articles, other error types emerged or increased. The number of words increased at each time, i.e., 93 words at Time1, 125 words at Time2, and 153 words at Time3. As Student 55 wrote more, this student also made more errors, i.e., eight errors at Time1, 11 errors at Time2, and 13 errors at Time3. However, the ratio of errors per 100 words were very similar at each time, i.e., 8.6 at Time1, 8.8 at Time2, and 8.5 at Time3. Next, I discuss what happened at each time.

At Time1, Student 55 made eight errors, one verb tense, two subject-verb agreements and five verb forms. An example of a verb form error pertained to infinitive verb form: "went" instead of "to go." Another example is the addition of "to" where it was not needed: "I think that excellent professors to help" instead of "help." An example of a subject-verb agreement error is an incorrect modification of plural form that should be singular: "he need students."

At Time2, Student 55 made 11 errors, four verb form, three preposition, two verb tense, and one each of subject-verb and article errors. It is worth noting that preposition and article

errors emerged for the first time at Time2. This student only slightly improved in subject-verb agreement. In terms of verb form errors, two of them are very similar to Time1, “bought” which should be an infinitive and “look”, which should be an infinitive or gerund. Another example is “we are listen” which should be “we are listening.” An example of a verb tense error is “he gives” which needs to be in the past. The rest of the specific sentence where this is located follows the past tense, so “gave” would maintain the verb tense continuity in that sentence. There were a few examples of preposition errors pertaining to time and location: “in 3:11 PM”. “in 1PM to 2PM” and “on 405 in the LPA room.” In terms of the one article error, “a excellent professor,” this student was correct to use an indefinite article but was not following the rule about “an” before a vowel sound.

At Time3, there was an increase in the number of verb form errors from Time2 to Time3, from four to eight errors, while the number of verb tense and preposition errors remained the same. The subject-verb agreement and article errors disappeared at Time3. There are two main commonalities among the types of verb form errors. One commonality is there are 5 missing infinitive verb forms, or misuses thereof, for example, the missing infinitive “He needs students online to bought” instead of “to buy.” Another commonality is the misuse of the gerund, for example, “we listening” that needs “are” before “listening” and “a professor needs to explaining” instead of “to explain.” In terms of preposition errors, the commonality is they are all related to date and time frame. For example, this student used two different prepositions in “at 3:50 – 4:450” and “in 1 to 2 PM.” “In” was also incorrectly used in “in Wednesday.” Yet, “in” was correctly used for location: “in 405 in the LPA”. The incorrect preposition use for time frames repeated itself from Time2 to Time3 while preposition usage for location improved from Time2 to Time3. In terms of verb tense error, an error type that repeated itself from Time2 to Time3 is the use of present tense “tell” where it should be past tense in order to be consistent with the rest of the sentence’s past verb tense usage.

To sum up for the three written performances in terms of accuracy for Student 55, this student had very similar scores for errors per 100 words at each time. As Student 55 wrote more words, the number of verb form errors increased, compared to the other error types at each time. According to various SLA researchers (e.g. Ellis, 2006; Lado, 1957; Odlin, 1989; Bardovi-Harlig, 2000), morphological complexity, of which verb form is an example, is very difficult for ESL learners whose L1 does not feature the same forms, i.e., L1 grammatical structures affect the way ESL students acquire verb morphology. The subject-verb agreement and article errors disappeared at Time 2. The preposition and verb tense errors remained the same at Times2 and 3.

Figure 5.4 showed Student 55's texts at all three performances in terms of sophisticated words. There were several similarities across the three writing samples. For example, Student 55 used the same number of sophisticated words, i.e., four sophisticated words at Times1 and 2. At Time2, there was a longer text, so slightly fewer, proportionally. Also, the sophisticated word "professor" appeared at least three times at each time. Next, I discuss what happened at each time in terms of sophisticated words. Afterwards, I identify several other complexity measures in terms of what simultaneously occurred.

At Time1, there were four sophisticated words out of 93 words, of which two of them were different from each other, i.e., professor and essays. At Time2, there were four sophisticated words out of 125 words, of which two of them were different from each other. However, none of the sophisticated words were new. In fact, at Times1 and 2, "professor" was used three times and "essay" was used once. At Time3, there were seven sophisticated words out of 153 words, (so proportionally similar to Time1) of which four were different from each other, and two were new. Two of the sophisticated words were repeated from Times1 and 2, i.e., professor and essay. Two new sophisticated words were dictionary and grammar.

Like with Student 10, specific measures of complexity followed slightly different patterns from one another for Student 55's texts. While lexical diversity remained the same at each time, average sentence length slightly decreased at each time. Simultaneously, Student 55's knowledge summary and knowledge transfer improved at Time2, then remained the same at Time3. Their knowledge summary score increased from three to four; knowledge transfer score increased from two to three.

Even though Student 55's written performance was overall the same across the repetitions, there was still enough of a scope for improvement. A possible reason why this student did not improve from the feedback is that the student might not have found it easy to learn from the feedback.

5.2.2 Knowledge summary and transfer & repetition

For **knowledge summary and transfer**, the analyses described in Chapter 4 (sections 4.2.1 and 4.2.2) revealed that, for both groups, there were statistically significant improvements with large effect sizes in knowledge summary and knowledge transfer at each time. As shown in Table 4-46, the effects of repetition went in the hypothesised directions, i.e., improvements with repetition at each time. This suggests that even with one task repetition of processing the listening input and performing the writing task, students already show improvements in their ability to reflect their understanding of the listening input into their writing, whether or not they receive feedback.

As the results from my study show, there were steady improvements in knowledge summary and knowledge transfer performances, and there was a main effect of repetition for both measures. This finding aligns with my hypothesis that written performance, like oral performance, improves with repetition and it extends to integrated writing. Regarding the fact that the listening aim in my study was to comprehend aural texts' main points, the task in my study had implications for general listening with a similar listening purpose (understanding the texts' main points) as well as more in-depth content from the listening input to then use in new contexts (knowledge transfer).

The findings from my study align with the findings from Sakai's (2009) and Iimura's (2006) task repetition studies on listening comprehension (**knowledge summary**), which consisted of listening-only tasks, introduced in section 2.7, in that significant improvements were made after repetition. In Sakai's study of 36 university EFL students in Japan, mainly in their second or third year, the findings revealed that both groups improved to a similar degree at Time2. As mentioned earlier, the students had studied English for six years before university study, and these two groups, i.e., higher listening proficiency group and lower listening proficiency group, were divided based on the mean score of the listening component of the Michigan Test. The large main effect of repetition and listening comprehension for both levels suggests that repetitions should continue in the classroom regardless of the proficiency level. Sakai's study, like my study, investigated these effects among university English students. While my study's participants were upper-intermediate level, results from Sakai's study that investigated two levels could inspire further investigations to compare knowledge summary performances across varying language proficiency levels to determine whether this would extend to all proficiencies.

In Iimura's (2006) study, the findings demonstrated that the mean scores for multiple-choice tests (Group A) were significantly higher at the first repetition (second performance), but no significant difference in scores between Time1 and 3. Mean scores for open-ended tests (Group B) were not significantly higher at the second performance, but the mean difference between Time1 and 3 was significantly higher. The same listening-only test was used repeatedly. Although Iimura's study compares two task types while my study uses one task, I refer to Iimura's study to (1) exemplify that repetition helps improve listening comprehension and (2) suggest that additional repetitions be incorporated for open-ended assignments.

The number of times a recording is played, the amount of time allotted between tasks, and the types of tasks can impact the results of students' performances on listening comprehension tests. This aligns with part of my study's design in that I played the recording

two times, and I provided one to two weeks, respectively, between task performances. Further investigation on task repetition studies that incorporate more varieties of open-ended tests would help confirm whether a similar finding on open-ended tests in Iimura's study would recur in future studies. In terms of **knowledge transfer**, I have not found any task repetition studies that investigated listening comprehension tasks that focused specifically on the integration of listening and writing or listening-reading-writing. To my knowledge, my study is the first one that has shown that repetition benefits knowledge transfer in an integrated listening-to-write task. However, as introduced in section 2.6, previous empirical studies on integrated tasks, though not task repetition studies, have used tasks that included skill integrations derived from major high-stakes examinations such as TOEFL (e.g., Plakans et al., 2019; Yang & Plakans, 2012), CAEL (Payant et al., 2019), and Pearson (PTE) (Rukthong, 2016). Following previous studies on integrated tasks, I incorporated an integrated task as opposed to an independent task to be in line with EAP settings where tasks with more than one language skill are typically used.

Next, I explain plausible reasons why both the *feedback* and *no-feedback* groups made similar continual significant improvements in knowledge summary and knowledge transfer at each time. One possible reason is that this was an exact repetition, so the listening input and the writing task were the same at each repetition, i.e., the students already knew exactly what they were being asked to write about. The students may have remembered some of the content from the input, at least in a more general way, so the repetition created a new opportunity for the students to listen to the input in an even more purposeful way. In this case, they could tweak their listening focus toward what they would have to write, i.e., they could listen to glean more details from the input to write about what they were being asked. Additionally, they may have previously processed the input, so they might have had more cognitive capacity to focus on more information and details from the input at repetitions and/or to think further about it. Hence, there would be more capacity for generating further content ideas to write about.

A possible reason that the feedback group did not outperform the no-feedback group in these respects might be related to the fact that the feedback consisted of scores and the broader rating scale descriptor statements. It did not specifically show exactly what in the piece of writing was strong versus what could be improved (this contrasts with the feedback method for the accuracy measures, which had shown exactly where in the performance errors had been made and what their nature was). In that sense, it is possible that had the feedback group been given very individualized, directed feedback on their knowledge summary and knowledge

transfer, they might have benefited from it more and outperformed the no-feedback group. As shown in Tables Table 4-42 and Table 4-44, the mean student scores from both groups increased to quite close to the maximum possible mean scores (in the upper band) at the last repetition for knowledge summary and knowledge transfer. However, because the mean scores were not at the maximum, the students as a group did not reach a ceiling effect. Perhaps, this means there could still be potential for a broader difference in written performances between the two groups had the feedback group received a larger scope of feedback, i.e., more precise and detailed.

5.3 Effect of feedback (RQ3)

RQ3 asked "Is there a change in listening-to-write task performance (CAF, knowledge summary and transfer) as a function of receiving feedback or not (Feedback)? If so, in which direction?"

The analyses in Chapter 4 revealed that there was no significant main effect of feedback for CAF measures. Similarly, there was no main effect of feedback for knowledge summary or knowledge transfer. As reported, the feedback group and the no-feedback groups made some significant improvements with repetition in various CAF measures. Where there were trade-off or partial trade-off effects in either group, accuracy was the strongest of the CAF dimensions versus complexity and fluency. In terms of knowledge summary and transfer, both groups made significant improvements, with somewhat similar mean increases at each time.

My findings for the lack of effect of feedback condition are opposite to some of the results from Nguyen et al.'s, (2023) oral task repetition study (introduced in section 2.4) in that there were effects of feedback in many of the findings in their study. In Nguyen et al.'s, (2023) study, the results showed that there were effects of feedback on some CAF measures for the task repetition group that received feedback. The teacher provided post-task teacher-corrected transcribing (corrective feedback). This group made significant improvements in subordination, verb forms, and speech rate (fluency). In this study, this feedback group made more significant improvements compared to the groups that did not receive feedback. The group that repeated the task but that did not receive feedback improved in mean length of units and verb forms; the control group (that did not repeat the same task or receive feedback) showed no significant improvements. What I can deduce from these findings from Nguyen et al.'s, (2023) study is that if I were to replicate my study, if I provide a wider range of feedback, it might result in significant effects of feedback.

Next is a previous study where my findings for the lack of effect of feedback somewhat align. Results from Jung's (2013) written task repetition study (introduced in section 2.4) suggest "the effect of feedback also seemed to be limited since all groups showed improvement regardless of the feedback condition" (p. 30). Due to the small number of participants in Jung's study, no statistical analyses were conducted. Meanwhile, some of the descriptive findings could shed some light. There were four groups: two of the groups received feedback, one that repeated the same task; one that completed a different task. The other two groups did not receive feedback: one repeated the same task; one completed a different task. The findings suggested improvements in four groups in their CAF performances. Most of the groups improved in terms of accuracy and fluency, with some slight decreases in complexity. However, findings in the groups based on feedback condition were not different from one another. Although Jung provided more detailed feedback than I did, Jung stated that a possible reason for this finding is the students were not given enough time to review the feedback. They were given 15 minutes to review their feedback before making the revisions. Unlike Jung's study, in my study, the students had about a one-week interval before their first repetition and about a two-week interval before their second repetition to reflect on the tasks that they completed despite not already being aware of when they would repeat the tasks.

The following is a study where the feedback delved into more detail compared to that of my study, and where semi-structured interviews were held. In Kim and Kim's (2017) study (introduced in section 2.4), they provided primarily indirect feedback (content) and some minimal direct feedback (error correction). The indirect feedback is where they provided open-ended comments such as asking for more details or suggesting better transitions in writing. The findings in the subsequent performances showed that there was continuous improvement except for the sixth performance, where some student performances remained the same while others became slightly worse. Findings in my study regarding the fact that there were some students who received the feedback but whose writing had not improved align with this aspect of Kim and Kim's findings. As the results from Kim and Kim's two student interviews suggest, they found the feedback helpful even though one of them did not improve in writing performance across repetitions. Also, there are lessons that I learned from Kim and Kim's study regarding how I could give feedback in future studies. For example, Kim and Kim provided more prompts and directions in student writing whereas in my study, the feedback that I provided for content were holistic scores along with their score band descriptors but without prompts in my students' written content.

Next, I discuss plausible reasons why there was no significant main effect of feedback for CAF, knowledge summary, or knowledge transfer. Also, I discuss possible reasons for some of the findings when I compared the feedback and no-feedback groups.

One possible reason is the kind of feedback that I gave. As introduced in the Methodology chapter, I gave feedback on areas that are known to be more challenging for these students, so I prioritized accuracy at the CAF component in the feedback. The linguistic measures that I selected are based on what I and the two EAP instructors I consulted with observed previously as a challenge for this population. I consciously decided to focus on just these aspects of the accuracy component of CAF (and the holistic scores for knowledge summary and transfer) because I did not want to overwhelm the students with too much feedback.

Here, I briefly explain why I selected knowledge summary and transfer as my other feedback points. In terms of knowledge summary, summarizing is an example of a skill that assesses listening comprehension in my integrated listening-to-write task. Recalling information addresses comprehension from the input. For an upper-intermediate level EAP class, a multiple-choice or true-false listening comprehension test format would not be in line with the level of difficulty of the course they were pursuing, hence why I had them write out their responses to the prompts. In terms of knowledge transfer, as this is an integrated listening-to-write task, I found it necessary to incorporate into the task, as introduced in section 3.5.2.2, some prompts where students start off applying their comprehension of the listening input, and then showing how they can use the information but through a prompt that related but did not ask for specific information that was specifically stated in the input. To keep in line with the level of rigor in this EAP course, my prompts required that the students applied their argumentative writing skills that required supporting details relevant to the input, thus an integration of listening and writing.

My conscious decision to not overwhelm them with feedback aligns with previous researchers' recommendations, as introduced in section 3.6, to refrain from providing excessive amounts of feedback, e.g., Lee (2013). My hypothesis was that all students would get better due to repetition, which the empirical data indeed support in many respects, but that the feedback group would get better to a larger extent on those elements they got feedback on as they would benefit from that feedback.

Next, I reflect further on the feedback that I gave in terms of how it played a role in the trade-offs. Since I gave feedback on the accuracy CAF dimension to the feedback group, this gave them an explicit insight into their accuracy errors, which meant they likely had increased

awareness regarding these and that it made it easier to address these in repetitions. In this case, the feedback group had a helpline for accuracy, on top of the repetition. They were guided on what the accuracy issues were.

Nevertheless, the no-feedback group also improved in accuracy with repetition (but less so). One possible way that improvements were possible for the no-feedback group is that there was exact repetition in the listening input. Possibly, the processing of the listening input became easier with repetition, and consequently, this gave the students more thinking time and/or cognitive capacity to focus on the linguistic aspects of their writing, such as improving the linguistic structures. This advantage of the same listening input, as introduced in section 5.2.2, was a chance for the students to use it in a more purposeful way to improve their writing, in this case not only for improving the content but also the grammar and structure. As the listening input and the writing prompts were the same, less of their thinking space needed to be devoted to processing the listening input. Another possible reason that improvements were possible for the no-feedback group is the students were in an EAP class. They were already in the frame of mind to improve the quality of their writing in various ways; therefore, they may have been geared toward writing better even without receiving feedback. The lack of feedback meant that they needed to self-identify accuracy issues and improve on them. They had to figure it out on their own, which is typically more challenging and they might not be aware of some issues in their own accuracy. This could explain why they improved in accuracy, but comparatively less so in complexity and fluency.

Additionally, the differing free cognitive capacities might also play an explanatory role in the trade-off results. Consequently, the differing free cognitive capacities might have given the students more space/time to think about the complexity and fluency aspects of their writing in the repeated performances. Therefore, it could mean that there was less attentional capacity left for the no-feedback group to focus on complexity and fluency, resulting in stronger trade-off's there even at the third performance, which might have taken more of their cognitive resources as compared to the feedback group, where there were partial trade-off's that disappeared at the third performance.

Next is some commentary on the potential for more elements of feedback for CAF in future studies. In this study, in terms of CAF, I focused my feedback on the accuracy measures, which were effective in that the feedback group had made more significant improvements in their accuracy measures earlier on in their repetitions than did the no-feedback group. As introduced in section 3.5.3.1, my selection of the specific linguistic targets for my accuracy feedback was based on information on students' written performances in the

EAP course outside of the research context to identify some strengths and weaknesses of student writing, i.e., based on my informal interviews with two instructors, along with my review of recent student essays completed soon before the start of my study. In hindsight, for future studies, I would extend the range of feedback to focus on lexis for both accuracy and complexity. An example would be sophisticated words, where I would recommend in the feedback that the student use more advanced vocabulary. Another example of lexis would be conjoining sentences. In this case, I would provide in my feedback examples of areas where short sentences could be combined to make longer and more meaningful sentences. In addition, I would recommend, where needed, for the student use of more transitional words. Further, as introduced in section 5.2.2, I would use more descriptive and individualized feedback for knowledge summary and knowledge transfer. For example, my feedback would be similar to that of Kim and Kim's (2017) study where they asked the students to explain more about what they meant by something that they wrote, provide more details, and where they made some suggestions like better transitions in their writing. With such areas of improvement in the range of feedback, potentially, this might result in a significant main effect of feedback on CAF and knowledge summary and transfer in future studies.

5.4 Student perceptions (RQ4a & RQ4b)

RQ4a asked "What are students' **perceptions** of the task used in this study, of listening-to-write tasks more generally, and of the extent to which integrated task repetition helps EAP students develop their writing proficiency?"

As detailed in the Results chapter in section 4.3, the analyses focused on three categories of post-task perception questionnaire responses: student perceptions about this task, listening-to-write tasks more generally, and about task repetition. In terms of **student perceptions about this task**, the overall student views based on the results from the 10 relevant Likert-scale statements were favorable, i.e. M=39.30 out of a total possible 50 points confirmed the students' positive views (see Table 4-48 and Table 4-49). This finding suggests that they would willingly write further versions of their work by doing even more repetitions in general to improve their written performances. The majority of the students' responses to the 10 statements were either "Agree" or "Strongly agree." For each statement, some students selected "Neutral" while only minimal numbers of students selected "Strongly disagree" or "Disagree." Overall, there were minimal variations in levels of agreement on the statements with one major exception: there was more variation in the range of responses to the statement about it being boring to repeat this task, as opposed to the range of responses to the other statements.

In terms of **student perceptions about listening-to-write tasks more generally**, the overall student views based on the results from the two relevant Likert-scale statements were favorable, i.e. M=9 out of a total possible 10 points confirmed the students' positive views about the way the listening-to-write integration helps improve writing skills (see Table 4-51 and Table 4-51). The majority of the students' responses to the 2 statements were either "Agree" or "Strongly agree", with very minimal "Neutral" responses selected. There were no "Strongly disagree" or "Disagree" responses selected. Therefore, there was no variation in the range of student responses. This suggests that students recognise that more than one skill can work in tandem to improve their language performance, in this case, listening and writing.

In terms of **student perceptions about task repetition**, data from student responses were based on four Likert-scale statements and one open-ended question about their perceptions in general about repeating a task. Based on the results from the Likert-scale statements (see Table 4-53 and Table 4-53), the majority of the students' views on repeating a task were favorable, i.e., M=17.89 out of a total possible 20 points confirmed their positive perceptions about repeating tasks. Based on the results from the open-ended question about students' overall opinions about task repetition ["What is your overall opinion of task repetition? Explain."], the majority of the students held favorable views about repeating a task. Out of the 64 students who participated in this study, 59 (92.2%) responded to this question. Out of the 59 students who responded to this question, 51 (86.4%) were positive opinions, 5 (8.5%) were negative opinions, and 3 (5.1%) were mixed opinions. Out of the 51 positive responses, 39 (76.5%) included explanations. Based on the content of these qualitative responses, the main categories of explanations were about listening skills, writing skills, a combination of both listening and writing, and general language skills. Thus, students offered multiple supportive reasons for the ways that repeating a task helps improve their language performances, thereby encouraging this practice to be used in future classrooms.

RQ4b asked "To what extent do student perceptions of task repetition differ between those who received feedback on their writing performances and those who did not?"

In addition to the overall student perceptions of this task and task repetition, as reported in detail in section 4.3, ANOVAs were run to determine whether there was a significant difference between the two groups' perceptions based on feedback condition. Results from these statistical tests revealed no significant difference between the two groups in terms of their general perceptions about this task, listening-to-write tasks more generally, or about

repeating a task in the future. These findings suggest that all students shared very positive views about this task and task repetition even if no feedback was provided.

As the results from my study show, the majority of the students held very favorable views about task repetition. This finding supports my hypothesis that students would recognise the benefits of task repetition. In fact, as there were no significant differences between the feedback condition groups in their views on task repetition, the benefits are recognised regardless of whether one receives feedback or not in addition to the repetitions.

The favorable views about task repetition found in the results of my study align with several other empirical studies, discussed in section 2.5, that investigated student perceptions about task repetition. The students in these studies were preparing for high-stakes examinations such as TOEFL and TOEIC. In Ahmadian et al.'s (2017) study, there were many examples of positive comments about task repetition, for example, "not boring" and "helpful" to more detailed responses such as ways that the repetition helped them improve their language production performances. Many of the open-ended responses my students wrote for the task repetition perception questionnaire align with some of the student responses in Ahmadian et al.'s study. Ahmadian et al. went a step further to capture the teachers' perceptions about their views on task repetition as well as what they anticipate would be their students' views. I did not explore the EAP instructor's views in my study, as I conducted the task administration myself without their involvement, but further studies that probe into this additional angle would help identify the degree of similarity between teachers' and students' views about task repetition. Such information could help guide teachers to determine how they would incorporate task repetition as they assess potential benefits as well as challenges for the students.

Another empirical study that investigated student perceptions about task repetition is Hanzawa and Suzuki's (2023) study. In Hanzawa and Suzuki's study, three groups of participants, grouped separately based on the amount of time interval between repetitions, were asked to complete a task repetition perception questionnaire. The majority of students held favorable views. Some of the statements that students selected included "I would do this task again", "I was bored doing this task", and "This practice excited my curiosity" (Hanzawa & Suzuki, p. 23). The results in the task perception part of my study align with those of Hanzawa & Suzuki. Also, there were no significant differences between the groups on task repetition perception in both Hanzawa and Suzuki's study and my study, thus a suggestion that task repetition should be incorporated regardless of the amount of time intervals that students are given between task performances. The findings in my study regarding there not being

significant differences between the groups' perceptions of task repetition also align with these additional findings in Hanzawa & Suzuki's (2023) study. A potentially useful feature of Hanzawa and Suzuki's perception investigation, which I did not look into in my study, is that they asked for views on the number of repetitions. In future studies, it could inform researchers and educators if more studies follow Hanzawa and Suzuki by adding a task repetition perception question that probes into the number of repetitions that the students feel would be enough for them to make improvements in their language performances. Such data of this sort would give educators a feel for an appropriate balance of engagement and performance, i.e., gathering such student views would help educators identify a sensible amount of repetition such that the students would not lose motivation to stay engaged with the tasks.

In sum, there have been several studies that reveal that students hold very favorable views about task repetition, and my study has reconfirmed this and extended it to the repetition of an integrated task type. Also, there have not been significant differences between different groups' perceptions in prior studies nor in my study, thus suggesting that regardless of whether they had enough time between repetitions (e.g. Hanzawa & Suzuki, 2023), received feedback, or were in a different level course (e.g. Norouzian et al., 2023; Ahmadian et al., 2017), almost everyone in the study was satisfied by doing task repetition.

5.5 Chapter summary

In this chapter, I presented a summary of the results of my analysis of the effects of task repetition and feedback on CAF and knowledge summary and transfer. I organized the placement of these summaries by research questions. I also stated whether the findings in my study aligned with my hypotheses. I then provided a discussion of the findings according to the key topic that most closely relates to each analysis, i.e., CAF and task repetition, effect of feedback on written performances, knowledge summary and knowledge transfer, and student perceptions about task repetition. I also discussed where there were trade-off effects among the CAF dimensions. Then I made connections between my findings and those in previous empirical studies. Where the findings in my study were not directionally consistent with the hypotheses, I provided possible reasons for those results.

6 Conclusion

6.1 Summary of the key findings

The aim of this study was to investigate the effect that task repetition has on student performances in an integrated listening-to-write task in terms of CAF, knowledge summary, and knowledge transfer. Feedback, an important pedagogic tool, was also examined. To that end, this study also examined the effect of feedback on students' academic writing performances. In addition, this study explored students' perceptions of this kind of task and of task repetition.

Findings from this study show that students can make significant improvements in terms of CAF, knowledge summary, and knowledge transfer in an integrated listening-to-write task by repeating the same task. This interpretation is supported by the fact that the statistical analyses revealed that there were significant main effects of task repetition for the majority of the CAF measures as well as for knowledge summary and knowledge transfer (listening comprehension) across three written performances. The majority of the CAF measures where there were main effects of repetition were the accuracy dimension, both global and by error-type. At the second performance, the analyses showed some competition among the CAF dimensions, i.e., accuracy competing most strongly with several complexity and fluency measures – which suggests a trade-off effect. More specifically in the second performance for the feedback group, a potential partial trade-off effect occurred between accuracy and several complexity measures that did not significantly improve; for the no-feedback group, the findings suggest a stronger trade-off occurred between accuracy and several complexity and fluency measures that significantly declined. In the third performance, for the feedback group, the finding suggest the trade-off effects disappeared; for the no-feedback group, trade-off effects appeared to remain

The findings in this study partially align with Skehan's (1998a, 1998b) Trade-Off Hypothesis in that at the second performance, there was competition between the CAF measures, though, unlike in previous studies, accuracy was stronger than complexity rather than the opposite. In addition to this, for the feedback group, most of the CAF measures' means that declined at the second performance improved at the third performance, i.e., the trade-off effects disappeared. This latter finding for the feedback group aligns with Sample and Michel's (2014) finding that as some dimensions benefit, they are in competition with other dimensions but that all dimensions improve by the third performance. This finding also

suggests that task repetition offers students many opportunities to progress in their language production.

The analyses in this study revealed that there were significant differences between the two groups' mean CAF scores at Time1. However, the students had been allocated randomly to their groups based on feedback conditions, so there is no clear explanation why there were some differences. There was, however, no main effect of feedback for the groups. There was, however, an interaction between feedback and repetition for some of the CAF measures, but not for knowledge summary or knowledge transfer. The majority of the CAF measures where there was an interaction were complexity and fluency, with only one accuracy measure by error type.

Overall, the findings suggest that task repetition and feedback are likely to support improvements in language performance, though these improvements varied across CAF measures. Several outcomes appeared to trend toward linear, upward directions. It should be acknowledged, nevertheless, that others appeared to trend toward non-linear, plateaued, or mixed patterns, notably for the no-feedback group. There were non-significant results that showed small to moderate effect sizes, which I reported for transparency. However, such findings should be interpreted with caution as potential trends rather than robust evidence (Plonsky & Oswald, 2014).

The majority of the students' views about task repetition, taken together, were very favorable. Similarly, the majority of the students had positive views about this task as well as integrated listening-to-write tasks. The analyses revealed that there were also no significant differences between the groups' perceptions concerning their views about this task, listening-to-write tasks or repeating tasks. A conclusion that I can draw from this finding is that most of the students would like to do more integrated tasks along with opportunities to repeat them to improve their language learning even if they do not receive feedback.

6.2 Contributions of the study

6.2.1 Theoretical contributions

This study provides important insights into Skehan's (1998a, 1998b) Trade-Off Hypothesis within the existing theory of TBLT. Within this existing theory, when a language learner is performing a task for the first time, the learner's capacity to simultaneously process the task demands and produce proper language is limited. At the second performance, some of the CAF dimensions compete with one another as the learner has a clearer understanding of the task and thus can allocate some more attention to the writing itself, but struggles to do this for all

dimensions equally. A trade-off usually occurs when a significant improvement in some CAF dimensions simultaneously accompanies significant declines in other CAF measures. Sometimes, the significant improvements are accompanied by a lack of significant gain or increase in other CAF measures.

Most of the body of research that has investigated task repetition effects in terms of CAF has been for oral performance. Only a small body of research has investigated repetition effects in terms of CAF in writing. The skill integrations under scrutiny have mainly been listening-to-speaking, reading-to-writing, and some reading-listening-write. Findings from this study extend the trade-off hypothesis to include integrated listening-to-write task repetition. In my study, there was competition among the CAF dimensions in the first and second performances. Only some of the CAF dimensions benefited from the repetition at the second performance, mainly accuracy as the strongest CAF dimension versus some measures of complexity and fluency. This finding from my study partially aligns with findings from earlier studies that investigated the impact of task repetition in oral language studies in that there was the presence of a trade-off. However, unlike previous studies, findings from my study showed accuracy as a stronger dimension to complexity rather than the opposite.

Most task repetition research and trade-off research has looked into one repetition (two performances in total). My study, however, extended this and explored the effect of an additional repetition, so two repetitions (three performances). In this regard, another contribution to knowledge derives from my study – at the third performance, for the feedback group, most of the trade-off effects have disappeared. This additional finding supports the application of Sample and Michel's (2014) oral task, two-repetition study and its finding that trade-off effects disappear at the third performance, and extends it to the context of integrated-writing task repetition. However, in my study, there were still trade-offs, although fewer than at their second performance, for the no-feedback group at the third performance. This suggests that without feedback, trade-offs might continue to some extent. Overall, the findings indicate that it may be worthwhile to allow students to do more than one repetition (and ideally also to give feedback).

6.2.2 Pedagogical contributions and implications

This study shows the importance of listening-to-write task repetition and incorporation of feedback in language learning. In terms of repeating a listening-to-write task, the majority of the average means of the CAF measures as well as knowledge summary and transfer significantly improved. While many of the CAF measures improved from Time1 to Time2,

some further improvements occurred when comparing Time1 to Time3. This finding suggests that even though some improvements occurred at one repetition, having the students repeat the task an additional time strengthens their improved language performances. Also, knowledge summary and knowledge transfer improved at each time, which suggests that learners make improvements in completing their tasks even if there is time for only one repetition. For this to be considered a task, it must always be meaning-based rather than form-focused even though grammatical features can be assessed as part of the task.

The use of real-time listening and then incorporating it into writing is an example of an authentic task demand used in academic, social and professional settings, thus an example of the meaning-based definition of an integrated task. Based on the findings in my study, the majority of the students responded on the post-task repetition perception questionnaire that they agreed about the extent to which listening for the purpose of writing helps develop language proficiency. Therefore, educators and researchers should continue to analyse the results when they integrate listening-to-write tasks as a basis to test the impact of task repetition to help build language proficiency.

Another advantageous method to incorporate is the use of integrated tasks. Tasks that integrate multiple language skills are instrumental in preparing students to achieve communicative goals necessary for academia and the real world, for example, taking notes, solving problems, simulation, etc. Such integration of skills provides more practice of real-life language use outside of classroom and testing settings. Integrated tasks are used in many high-stakes language examinations; thus this tool helps prepare students to succeed in multiple ways. Providing many listening tasks entices learners to construct meaning from the input, and by integrating it into writing, this integration allows students more time to reflect on and view their output, for example, making error corrections as they practice writing, than they would if the task were a listening-to-speak integration. Students in my study expressed favorable opinions about the extent to which task repetition helped them improve language learning. It is possible to conclude that integrated task repetition could be incorporated with listening and writing as well as other skill integrations that include at least one of these two skills.

Regarding feedback, my study's findings showed no main effect for the feedback condition in isolation. Potentially, it is not enough to provide merely coded metalinguistic feedback and holistic scores, i.e. accuracy - the nature of the CAF feedback in my study, to result in significant main effects in isolation. However, the interaction between feedback condition and repetition shows that by offering feedback together with repetitions, there are pedagogic effects. A contribution to knowledge that this finding suggests is that feedback is

worthwhile, but there should be extended ranges of feedback beyond those offered in my study in order for feedback in isolation to produce significant effects.

An additional contribution from this study to pedagogy worth noting is that even if feedback is not provided, it might not impact students' perceptions about this task or task repetition. Findings in my study showed no significant differences between the feedback group and no-feedback groups in terms of their favorable views about this task and desire to repeat a task. Combined with the finding of a main effect of repetition for many measures, this suggests that it is worthwhile for educators to incorporate task repetition regardless of the type of feedback or even if they do not have the resources to offer feedback.

In sum, from a pedagogical angle, the findings in this study do not show any disadvantages to student learning from integrated task repetition; crucially, the findings suggest that repeating a task (ideally accompanied by feedback) offers many advantages to faculty and students in EAP programs for improving students' language proficiency.

6.3 Limitations and further research

In this section, I explain limitations of the present study, which concern shortcomings in the task's design, overlap between the definitions of the knowledge summary and knowledge transfer constructs, and the type of CAF feedback provided to the students. Then, I provide some recommendations to broaden the participant pool and the post-task repetition perception questionnaire.

6.3.1 Shortcomings of the task used

6.3.1.1 Task authenticity (pedagogical vs. real-life use)

The integrated task used in my study aligns with pedagogical practices in the EAP classroom, although there are a few limitations. In favour of the task, the listening-to-write skill integration in this task, unlike reading-to-write, where there are opportunities to reread structured written text input, involves careful processing of information in real-time that is often equivalent to unstructured input. This skill integration is uniquely relevant because it requires the ability to process key information, which involves specific cognitive and linguistic challenges to capture the input. An advantage of listening-to-write, unlike listening-to-speak, is it allows learners to review what they write as they produce the language. The task in this study was to look at the ability to process real-time aural information, which-could then be produced in writing. However, the task might not fully reflect authentic academic, social, or professional practices that learners will use outside of EAP settings. For example, in real life, summarization tasks typically draw on multiple as well as more complex sources (e.g.,

audiences with more directed communicative goals, a combination of inputs, both visual and aural, etc.). This study's task design was conducted in an EAP classroom setting, i.e., pedagogical instead of authentic. The task used in this study, however, was appropriate for the EAP setting, i.e., it did represent the pedagogic task of the EAP classroom where I conducted the research. Conducting this study in a controlled classroom setting helped ensure comparability between the two participant groups. However, the study's findings might not thoroughly extend to everyday contexts outside the EAP classroom. This delicate balance between experimental control and real-life connections is a challenge in classroom-based research.

Building on the balance between experiment and real-life connections, the presence of many academic and professional multimodal settings that combine visual and aural input may suggest that an aural-only input in a listening-to-write task may not be authentic in other academic domains or in real life. In future research, adding more multimodal content to the input, such as a video of a lecturer introducing the class while using visuals such as slides and other images, would help authenticate this task.

Also, the real-time aural information in this study was limited to only one integrated task that included the same listening input and writing prompt, i.e. exact task repetition. In this case, procedural repetition was not the focus, so it remains unclear whether similar positive findings would occur if different listening-to-write tasks (or at least different inputs and/or outputs) had been used at each time.

6.3.1.2 Absence of word-count control (summary vs. paraphrase ambiguity)

The absence of a word-count requirement for the summary task (Task Question 1) is another potential limitation. While this decision gave the participants more flexibility in the amount of writing they produced, it may have somewhat blurred the boundary between summarizing and paraphrasing. In this case, some responses may resemble paraphrases with similar original word-count rather than concise summaries. However, researchers (e.g., Hirvela & Du, 2013; Hyland, 2004; Keck, 2006) have shown that it is common for there to be an overlap of summarizing and paraphrasing in academic work in higher education classrooms, thereby suggesting that this blurring reflects authentic student writing. Paraphrasing is commonly practiced in summary writing, and this tends to blur the distinction between summarizing and paraphrasing.

Although the absence of word-count control is a potential limitation, an informal analysis of the discourse of the written performances in my study does not suggest major shifts to paraphrasing. Also, in my study, the focus of my task was not to test ability to write a

condensed summary but rather to examine the way students integrate listening to writing skills such that they could process, retell, and use the information in another context. In this case, both summarizing and paraphrasing align with writing based on input.

6.3.1.3 Third task question not fully aligned with listening input

Another potential limitation concerns the third prompt question of the listening-to-write task, "Discuss at least three (3) qualities you look for in an excellent professor. Explain why." The instruction is not directly tied to the information in the listening input and is, therefore, a shift away from summarization. As introduced in section 3.5.2.2, it was not fully the idea that this question would be a summary because this question was attempting to elicit evidence of knowledge transfer. At the same time, there is a risk of knowledge transfer in that a learner might at times rely less on content from the listening input and, instead, draw more on prior knowledge or personal experiences**.** As a result, the response could be less anchored to the input. This task question was related to student success, which required students to think beyond the words from the input rather than use specific content from the input; thus, the output cannot be attributed solely to summarization. Even so, the underpinning topic for the third question of the task, though not directly tied to the input itself, still pertained to college student success, which relates to some components of the input. In a future study, explicitly requiring students to reference information from the input could strengthen the alignment with knowledge transfer.

6.3.2 Boundary Between Knowledge Summary & Transfer

The overlap between the distinct definitions of knowledge summary and knowledge transfer represents another limitation. In line with the literature on these constructs (see section 2.7) and with their operationalisation in rating scales (see section 3.5.3.2), I kept these constructs separate in my study. However, some of the participants' texts blended features of both. This overlap can invite debate on the ability of researchers to align written performance with distinct categories, thus potentially influencing the way performance is interpreted. The presence of this blur between the constructs suggests that the findings should be interpreted with caution. At the same time, this blur between the constructs is not unique to my study. As discussed in section 2.7, students are usually expected to simultaneously apply both summarizing and transfer skills in academic writing. Therefore, the participants' writing that combined both constructs goes beyond a limitation in the clarity of analytic features but reflects the practicality of the way academic writing is typically performed. Future researchers who explore written performance while teasing apart the two constructs may encounter a

challenge: they might need to factor into consideration the way that this blurred boundary could shape their interpretation of the findings.

6.3.3 CAF feedback focus only on accuracy

The type of feedback I provided to students is another limitation. In my study, CAF feedback was limited to accuracy, while other CAF dimensions, i.e., complexity and fluency (e.g., vocabulary, sentence structure, etc.), were not focused on in the feedback. It is possible that after the participants in the feedback group had been given the first feedback, they might have inferred that error correction was the main/exclusive focus of the task, influencing their attention during repetitions. It could be that they paid more attention to accuracy than to the other aspects of their writing. As presented in Table 4-38, the feedback group demonstrated fewer significant improvements in complexity and fluency from Time1 to Time2 and Time2 to Time3 compared to accuracy. These findings suggest that this group might have made further significant improvements in complexity and fluency had the CAF feedback extended beyond accuracy. Although these suggestions cannot be completely confirmed, the accuracy-only focus of feedback might have affected the participants' perception of the task toward grammatical accuracy rather than in broader aspects of their writing. Future studies might benefit by providing an extended range of feedback that covers more CAF components, thus providing a broader overview of how feedback and repetition interact to support language learner development in writing.

Next are two remaining limitations of this study in that there is room to broaden the participant pool and to enhance the post-task repetition perception questionnaire design.

6.3.4 Broadening the participant pool

An additional limitation of my study pertained to the participant pool. My study included upper-intermediate students in a university EAP class. However, what this study did not explore was a comparison in listening-to-write task repetition performance among different proficiency groups, e.g. intermediate, upper-intermediate, and advanced. A comparative study among several proficiency-level classes using level-appropriate tasks, yet following the same procedures for all levels in the study, might shed light on whether the findings from my study could generalize across proficiency levels. An additional comparison could involve expanded participant groups. For example, there could be four groups: a feedback group with exact task repetition; a feedback group with procedural repetition (a different task) ; a no-feedback group with exact task repetition; and a no-feedback group with procedural repetition.

Moreover, adding variability to the participant structure, as well as increasing the number of participants in future studies, could help address the earlier point about the non-significant results with small or moderate effect sizes. Future research with larger numbers of participants or a more varied participant structure would help determine whether such patterns from this study represent robust trends (Plonsky & Oswald, 2014).

6.3.5 Enhancing the post-task repetition perception questionnaire

Another limitation relates to the fact that most of my post-task repetition questionnaires were quantitative with only one open-ended question to further capture the students' opinions about task repetition. A semi-structured interview might have gathered even more, in-depth opinions that students might have shared about this task, the listening-to-write integration or repeating tasks. Even so, in this study, there was a wide range of detail that the participants provided in their responses to the open-ended question. Also, the questionnaire did not include a statement or question to probe into student views on usefulness of feedback and the kind of feedback that they feel helps their learning the most. Such questionnaire design expansion should capture these perceptions more fully. Responses to such questions might have helped me further discuss possible reasons why there was no main effect between the feedback condition groups.

In sum, the above limitations suggest that the findings from the present study should be interpreted with reference to the study's methodological characteristics. The results do offer useful insights into task repetition, knowledge summary and transfer, and feedback. However, the findings are also shaped by the choices made in the design of this study. Acknowledging these limitations may help inform future empirical research that aims to investigate with more task authenticity, wider ranges of feedback, and more diverse student populations. In turn, such studies could potentially provide additional insight into integrated listening-to-write performances.

References

Abbasian, G. & Chenabi, F. (2016). Language Skill-Task Corollary: The Effect of Decision-Making vs. Jigsaw Tasks on Developing EFL Learners' Listening and Speaking Abilities. *The Journal of Applied Linguistics. (Online), 9*(18), 1-24.

Abdelaty, S. (2023). Overcoming Obstacles to TBLT and CLT Implementation: A Study of Libyan English Language Teachers' Perspectives and Strategies. *International Journal of Scientific and Research*, *13*(4), 238-245. https://doi.org/10.29322/IJSRP.13.04.2023.p13632

Abrams, Z. (2019). The effects of integrated writing on linguistic complexity in L2 writing and Task complexity. *System (Linköping)*, *81*, 110–121. https://doi.org/10.1016/j.system.2019.01.009

Ahmadian, M. J. (2011). The effect of 'massed' task repetitions on complexity, accuracy, and fluency: does it transfer to a new task? *The Language Learning Journal, 39,* 269–280.

Ahmadian, M. J. (2012). The effects of guided careful online planning on complexity, accuracy and fluency in intermediate EFL learners' oral production: The case of English articles. *Language Teaching Research*, *16*(1), 129-149. https://doi.org/10.1177/1362168811425433

Ahmadian, M.J. and Tavakoli, M. (2011) The effects of simultaneous use of careful online planning and task repetition on accuracy, fluency, and complexity of EFL learners" oral production. *Language Teaching Research, 15(*1), 35-59, doi:10.1177/1362168810383329

Ahmadian, M. J., Mansouri, S. A., & Ghominejad, S. (2017). Language learners' and teachers' perceptions of task repetition. *ELT Journal, 71,* 467–477. https://doi.org/10.1093/elt/ccx011

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge University Press.

Allman, B. (2019). *Effective and Appropriate Feedback for English Learners.* In B. Allman (Ed.), Principles of Language Acquisition. EdTech Books. Retrieved from https://edtechbooks.org/language_acquisition/effective_el l_appropriate_feedback

Amiryousefi, M. (2016) The differential effects of two types of task repetition on the complexity, accuracy, and fluency in computer-mediated L2 written production: A focus on computer anxiety, *Computer Assisted Language Learning, 29*(5), 1052-1068, DOI:10.1080/09588221.2016.1170040

Azizzadeh, L., & Dobakhti, L. (2015). The Effect of Task Repetition on Complexity and Accuracy of Iranian High-intermediate EFL Learners' Narrative Writing Performance. *International Journal of Applied Linguistics and English Literature, 4*, 17-25.

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing*, *19*(4), 453–476. https://doi.org/10.1191/ 0265532202lt240oa

Bardovi-Harlig, K., & Bofman, T. (1989). Attainment of syntactic and morphological accuracy by advanced language learners. *Studies in Second Language Acquisition* , *11,* 17-34.

Bardovi-Harlig, K. (2000). *Tense and aspect in second language acquisition: Form, meaning, and use*. Wiley-Blackwell.

Barrot, J.S., & Agdeppa, J.Y. (2021). Complexity, accuracy, and fluency as indices of college-level L2 writers' proficiency. *Assessing Writing, 47*, 100510.

Bayat, N. (2018). A Comparative Study of the Effects of Task Repetition, Unguided Strategic Planning, and Pressured On-line Planning on the Accuracy of Upper-intermediate EFL Learners' Written Production. *Linguistics and Literature Studies, 6*, 1-11.

Bellon, J., Bellon, E., & Blank, M. A. (1991). *Teaching from a research knowledge base: A development and renewal process*. Merrill.

Benavent, G. T., Penamaría, S. (2011). Use of authentic materials in the ESP classroom. *Encuentro 20,* 89-94.

Biber, D., & Gray, B. (2013). *Discourse characteristics of writing and speaking task types on the TOEFL iBT Test: A lexico-grammatical analysis.* (TOEFL iBT Research Report RR-19). Princeton, NJ: Educational Testing Service.

Boers, F. (2014). A reappraisal of the 4/3/2 activity. *RELC Journal, 45*(3), 221–235. https://doi.org/10.1177/0033688214546964

Brezina, V. & Pallotti, G. (2016). Morphological complexity in written ESL texts. *Second Language Research,* Advance Access [OPEN ACCESS].

Brindley, G. (1994). Task-centred assessment in language learning programs: the promise and the challenge. In N. Bird, P. Falvey, A. Tsui, D. Allison, & A. McNeill (Eds.), *Language and learning* (pp. 73–94). Hong Kong: Institute of Language in Education.

Brindley, G. (2013). TBLA. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1-6). Wiley Blackwell. https://doi.org/10.1002/9781405198431

Brown, J. E., (2001) "Learning through Listening Strategies for Literature," *Language Arts Journal of Michigan*, *(17)*2. Available at: https://doi.org/10.9707/2168-149X.1316

Brown, H. (2007). Teaching by Principles: *An Interactive Approach to Language Pedagogy (3rd ed.).* London: Pearson Education ESL.

Bui, G., Ahmadian, M. J., & Hunter, A.-M. (2019). Spacing effects on repeated L2 task performance. System, 81, 1–13. https://doi.org/10.1016/j.system.2018.12.006

Bulté, B. and Housen, A. (2012) Defining and Operationalizing L2 Complexity. In: Housen, A., Kuiken, F. and Vedder, I., Eds., Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA, John Benjamins, Amsterdam, 21-46. https://doi.org/10.1075/lllt.32.02bul

Bulté, B., Housen, A., & Pallotti, G. (2024). Complexity and difficulty in second language acquisition: A theoretical and methodological overview. *Language Learning*, 1-42. https://doi.org/10.1111/lang.12669

Bulté, B., Housen, A., Pierrard, M., & Van Daele, S. (2008). Investigating lexical proficiency development over time - the case of Dutch-speaking learners of French in Brussels. *Journal of French Language Studies, 18*(3), 277-298. doi: https://doi.org/10.1017/S0959269508003451

Butler, Y. (2011). The implementation of communicative and task-based teaching in the Asia-Pacific region. Annual Review of Applied Linguistics 31, 36-57.

Bygate, M. 2006. "Areas of research that influence L2 speaking instruction" in E. Uso´-Juan and A. Martinez-Flor (eds.). *Current trends in the development and teaching of the four skills.* Berlin: Mouton de Gruyter.

Bygate, M. (1996). Effects of task repetition: appraising the developing language of learners. In Willis, J., and Willis, D., (Eds.). Challenge and Change in Language Teaching. Oxford: Heinemann. 136-146.

Bygate, M. (2001). Effects of task repetition on the structure and control of oral language. In Bygate, M., Skehan, P., & Swain M. (Eds.), *Researching pedagogic tasks*. Harlow: Longman. 23-48

Bygate, M. (2009). Effects of task repetition on the structure and control of oral language. In K.Van den Branden, M. Bygate, & J. Norris (Eds.), *Task-based language teaching: A reader* 249–274. Philadelphia, PA: John Benjamins.

Bygate, M. (2018). (Ed.). *Learning language through task repetition.* John Benjamins Publishing Company

Bygate, M. (2016). Sources, developments and directions of task-based language teaching. *Language Learning Journal*, *44*(4), 381–400. https://doi.org/10.1080/09571736.2015.1039566

Bygate, M. (1999). Task as the context for the framing, re-framing and unframing of language. *System 27*: 33 – 48.

Bygate & Samuda (2005). Integrative planning through the use of task repetition. In Ellis, R. (ed.), *Planning and task performance in second language.* Amsterdam: John Benjamins. 37-74.

Caparas, P. (2022). Metalinguistic Corrective Feedback and Students' Response to Feedback in L2 Writing. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*, 682–690. Manila, Philippines

Chan, S., & May, L. (2023). Towards more valid scoring criteria for integrated reading-writing and listening-writing summary tasks. *Language Testing*, *40*(2), 410–439. https://doi.org/10.1177/02655322221135025

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). *Building a validity argument for the test of English as a foreign language.* Routledge.

Chodorow, M., Tetreault, J., & Han, N.-R. (2007). "Detection of grammatical errors involving prepositions." *Proceedings of the Fourth ACL-SIGSEM Workshop on Prepositions.*

Cobb, T. (2017). VOCABPROFILE ENGLISH, an adaptation of Heatley, Nation & Coxhead's (2002) *Range*. Retrieved July 17, 2019, from http://www.lextutor.ca/vp/eng/

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillside, NJ: Lawrence Erlbaum Associates.

Cohen, L., Manion, L., & Morrison, K. (2018). *Research methods in education (8th ed.).* Routledge.

Cooney, A., Darcy, E., & Casey, D. (2018). Integrating reading and writing: supporting students' writing from source. *Journal of University Teaching & Learning Practice, 15*(5). https://ro.uow.edu.au/jutlp/vol15/iss5/3.

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*. Council of Europe Publishing. https://www.coe.int/lang-cefr

Council of Europe (2023, January 1). *Global scale - Table 1 (CEFR 3.3): Common Reference levels*. Common European Framework of Reference for Languages (CEFR). Retrieved April 1, 2023, from https://www.coe.int/en/web/common-european-framework-reference-languages/table-1-cefr-3.3-common-reference-levels-global-scale

Crookes, G. 1989: Planning and interlanguage variation. *Studies in Second Language Acquisition*, *11,* 367-383 .

Cumming, A. (2014). *Assessing integrated skills.* In A. Kunnan (Ed.), Companion to language assessment (pp. 216e229). Malden, MA: Wiley-Blackwell. http:// dx.doi.org/10.1002/9781118411360.wbcla131.

Cumming, A. (2013). Assessing integrated writing tasks for academic purposes: Promises and perils. *Language Assessment Quarterly, 10*(1).

Cumming, A. (Ed.). (2006). Goals for academic writing: ESL students and their instructors. *Studies in Second Language Acquisition*, *30*(2), 271–271. doi:10.1017/S0272263108080443

Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., James, M. (2006). Analysis of discourse features and verification of scoring levels for independent and integrated tasks for the new TOEFL. (TOEFL Monograph No.MS-30 RM 05–13). Princeton, NJ: Educational Testing Service.

Cumming, A. H., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing, 10*(1), 5–43.

Cumming, A., Kantor, R. Powers, D., Santos, T., & Taylor, C. (2000). *TOEFL 2000 writing framework: A working paper.* TOEFL Monograph 18. Princeton, NJ: Educational Testing Service.

Davison, I. (2021). *The Effects of Carrying Out Collaborative Writing on the Individual Writing Proficiency of English Second Language Learners in an English for Academic Purposes Program*. Lancaster University (United Kingdom).

Davison, I. (2024). The effects of completing collaborative or independent writing on the development of language use in individual writing. *Journal of Second Language Writing*, *65*, 1-13. https://doi.org/10.1016/j.jslw.2024.101128

Diab, N. (2015). Effectiveness of written corrective feedback: Does type of error and type of Correction matter? *Assessing Writing, 24,* 16–34.

Domínguez, L., Arche, M. J., & Myles, F. (2017). Testing the Feature Reassembly Hypothesis: Tense and aspect in L2 Spanish. *Second Language Research, 33*(1), 67–100. https://doi.org/10.1177/0267658317701991

Dörnyei, Z. & Kormos, J. (1998). Problem-solving mechanisms in L2 communication: A psycholinguistic perspective. Studies in Second Language Acquisition, 20, 349–385.

East, M. (2012). *Task-based language teaching from the teachers' perspective.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

Education First (2022, January 1). *English level B2*. Upper Intermediate (EF SET Score 51-60). Retrieved April 4, 2023, from https://www.efset.org/cefr/b2/

Ellis, R. (2005). Planning and task-based performance: Theory and research. In Ellis, R. (ed.), Planning and task performance in second language. Amsterdam: John Benjamins. 3-34.

Ellis, R. (2018). Reflections on task-based language Teaching: St. Nicholas House.

Ellis, R. (2003). *Task-based language learning and teaching.* Oxford: Oxford University Press.

Ellis, R. (2000). Task-based research and language pedagogy. *Language Testing Research*, *4*(3), 193-200. 15

Ellis, R. (2009). The differential effects of three types of task planning on the fluency, complexity, and accuracy in L2 oral production. *Applied Linguistics*, *30*, 474-509.

Ellis, R. (2006). *The study of second language acquisition* (2nd ed.). Oxford University Press.

Ellis, R., & Shintani, N. (2013). *Exploring language pedagogy through second language acquisition research*. New York: Routledge.

Ellis, R., Skehan, P., Li, S., Shintani, N., Lambert, C. (2019). Task-Based Language Teaching: Theory and Practice. Cambridge University Press.

Evans, N. W., Hartshorn, K. J., Cox, T. L., & de Jel, T. M. (2014). Measuring written Linguistic accuracy with weighted clause ratios: A question of validity. *Journal of Second Language Writing, 24,* 33–50.

Faez, F., & Tavakoli, P. (2019). *Task-based Language Teaching*.: TESOL International Association

Fellner, T. & Apple, M. (2006). Developing writing fluency and lexical complexity with blogs. *JALT CALL Journal, 2*(1), 15-26.

Ferreira, M. M., Arroyo, A. T., & Druckman, M. F. (2007). Action research: Assessing student learning through the scientific method. *Science Scope*, *30*(8), 64–67.

Ferris, D. (1994). Lexical and syntactic features of ESL writing by students at different levels of L2 proficiency, *TESOL Quarterly, 28*, 414–420.

Ferris, D. (2010). Second language writing research and written corrective feedback in SLA: Intersections and practical applications. *Studies in Second Language Acquisition 32*(2), 181–201.

Ferris, D. (1999). The Case for Grammar Correction in L2 Writing Classes: A Reponse to Truscott (1996). Journal of Second Language Writing, 8, 1-11. https://doi.org/10.1016/S1060-3743(99)80110-6

Ferris, D. R. (2002). *Treatment of error in second language student writing.* Ann Arbor, MI: The University of Michigan Press.

Ferris, D. R., & Hedgcock, J. S. (2014). *Teaching L2 composition: Purpose, process, and practice* (3rd ed.). Routledge.

Fleckenstein, J., Keller, S., Krüger, M., Tannenbaum, R. J., & Köller, O. (2020). Linking TOEFL iBT® writing rubrics to CEFR levels: Cut scores and validity evidence from a standard setting study. *Assessing Writing*, *43*, 33-47. https://doi.org/10.1016/j.asw.2019.100420

Foster, P., & Skehan, P. (1996). The influence of planning on performance in task-based learning. *Studies in Second Language Acquisition, 18*(3), 299-324.

Freed, B. 2000. Is fluency, like beauty, in the eyes (and ears) of the beholder? In Heidi Riggenbach (ed.), Perspectives on Fluency, 243–265. Ann Arbor: University of Michigan Press.

Fukunaga. (2021). L2 writing development through two types of writing task repetition. *International Review of Applied Linguistics in Language Teaching, IRAL*. https://doi.org/10.1515/iral-2021-0144

Fukuta, J. (2016). Effects of task repetition on learners' attention orientation in L2 oral

production. *Language Teaching Research : LTR*, *20*(3), 321–340. https://doi.org/10.1177/1362168815570142

Gass, S. & Selinker, L. (2001). *Second language acquisition: An introductory course.* Lawrence Erlbaum

Gass, S., Mackey, A., Alvarez-Torres, M. J., & Fernández-García, M. (1999). The effects of task repetition on linguistic output. Language Learning, 49, 549-581.

Gebril, A. (2006). Independent and integrated academic writing tasks: a study in generalizability and test method. Unpublished doctoral dissertation. The University of Iowa.

Gebril, A. & Plakans, L. (2008). *Investigation of source use, discourse features, and process in integrated writing tests.* Spaan Working Papers.

Gebril, A. & Plakans, L. (2013). Toward a transparent construct of reading-to-write tasks: The relationship between discourse features and proficiency. *Language Assessment Quarterly, 10*, 9–27.

Geiwitz, P. J. (1966). Structure of boredom. *Journal of Personality and Social Psychology, 3,* 592–600. *https://psycnet.apa.org/doi/10.1037/h0023202*

Ghahderijani, B. H. (2021). The impact of of corrective feedback on Iranian EFL learners' writing complexity, accuracy, and fluency. *Journal of Nusantara Studies (JONUS)*, *6*(2), 250-272. https://doi.org/10.24200/jonus.vol6iss2pp250-272

Grabe, W., & Zhang, C. (2013). Reading and writing together: A critical component of English for academic purposes teaching and learning. *TESOL Journal, 4*(1), 9–24. https://doi.org/10.1002/tesj.65Green, R. (2013). *Statistical analyses for language testers.* Palgrave Macmillan.

Hamp-Lyons, L., & Henning, G. (1991). Communicative writing profiles: An investigation of the transferability of a multiple-trait scoring instrument across ESL writing assessment contexts. *Language Learning, 41,* 337–373.

Hanzawa, K., & Suzuki, Y. (2023). How do learners perceive task repetition? Distributed practice effects on engagement and metacognitive judgment. *Modern Language Journal, 107,* 1–28. https://doi.org/10.1111/modl.12843

Hassanzadeh-Taleshi, M., Yaqubi, B., & Bozorgian, H. (2023). The effects of combining task repetition with immediate post-task transcribing on L2 learners' oral narratives. *Language Learning Journal*, *51*(2), 133–144. https://doi.org/10.1080/09571736.2021.1901967

Haskell, E. (2001). *Transfer of learning: Cognition, instruction and reasoning.* New York: Academic Press.

Hawkes, M. L. (2009). Effects of task repetition on learner motivation. In A.M. Stoke, *JALT2009 Conference Proceedings.* Tokyo: JALT.

Hawkes, M. L. (2012). Using task repetition to direct learner attention and focus on form. *ELT Journal,66*(3), 327–336.

Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In Levy, C.M. & Ransdell, S. E. (Eds.), *The science of writing*. Mahwah, NJ: Erlbaum. 1-27.

Heatley, A., Nation, I., & Coxhead, A. Range and frequency programs. Retrieved March 1, 2023 from: http://www.victoria.ac.nz/lals/staff/paul-nation.aspx

Hidi, S. & Anderson, V. (1986). Producing written summaries: Task demands, cognitive Operations, and implications for instruction. *Review of Educational Research. 56*(4), 473 – 93.

Hinkel, E. (2006). Current perspectives on teaching the four skills. *TESOL Quarterly, 40*(1), 109-131

Hirvela, A., & Du, Q. (2013). "Why am I paraphrasing?": Undergraduate ESL writers' engagement with source-based academic writing and reading. *Journal of English for Academic Purposes, 12*(2), 87–98.

Holger, D. (2004). *The Acquisition of Complex Sentences.* Cambridge University Press.

House, S. 2008. «Authentic materials in the classroom». *In Didactic approaches for teachers of English in an international context.* Sonsoles Sánchez-Reyes Peñamaría and Ramiro Durán Martínez, 53-70. Salamanca: Ediciones Universidad de Salamanca.

Housen, A., & Kuiken, F. (2009). Complexity, accuracy and fluency in second language acquisition. *Applied Linguistics, 30*(4), 461-473.

Hsu, H. C. 2019. The combined effect of task repetition and post-task transcribing on L2 speaking complexity, accuracy, and fluency. *The Language Learning Journal,* 47(2)*,* 172–87. doi:10.1080/09571736.2016.1255773.

Hunt, K. (1965). *Grammatical structures written at three grade levels.* National Council of Teachers of English.

Iimura, H. (2006). The effects of task difference and repetition in listening comprehension. *Japan Society of English Language Education*, *17*, 51-60. https://doi.org/10.20581/arele.17.0_51

Indrarathne, B. (2013). *Effects of task repetition on written language production in task-based language teaching.* Paper presented at the Lancaster University postgraduate conference in linguistics & language teaching. Lancaster, UK.

Ionin, T., & Wexler, K. (2003). The certain uses of *the* in L2-English. In J. M. Liceras, H. Zobl, & H. Goodluck (Eds.), *Proceedings of the 6th Generative Approaches to Second Language Acquisition Conference (GASLA 6)* (pp. 150–160). Cascadilla Press.

Ionin, T., & Wexler, K. (2002). Why is 'is' easier than '-s'? Acquisition of tense/agreement

morphology by child second language learners of English. *Second Language Research, 18*(2), 95–136. https://doi.org/10.1191/0267658302sr195oa

Irmawan, N., & Nurdini, R. A. (2020). The errors of EFL students' TOEFL iBT integrated writing task. *Journal of Research on English and Language Learning (J-REaLL)*, *1*(2), 73–86. https://doi.org/10.33474/j-reall.v1i2.6860

Johns, M. (1985). Summary protocols of "unprepared" and "adept" university students: Replications and distortions of the original. Language Learning and Technology, 14(3), 51 – 71.

Jung, S. (2013). The effect of task repetition and corrective feedback in L2 writing: a pilot study. MSU Working Papers in SLS, 4, 24-38.

Jung, Y., Kim, Y., & Murphy, J. (2017). The role of task repetition in learning word-stress patterns through auditory priming tasks. *Studies in Second Language Acquisition*, *39*(2), 319–346. https://doi.org/10.1017/S0272263117000031

Keck, C. (2014). Copying, paraphrasing, and academic writing development: A re-examination of L1 and L2 summarization practices. *Journal of Second Language Writing, 25*, 4–22.

Kellerman, E. (1983). Now you see it, now you don't. In S. Gass & L. Selinker (Eds.), *Language transfer in language learning* (pp. 112–134). Newbury House.

Kellogg, R. T. (1996). A model of working memory in writing. In Levy C.M. & Ransdell, S.E. (Eds.), *The science of writing.* Mahwah, NJ: Erlbaum. 57-71.

Kellogg, R. T. (2001). Commentary on processing modalities and development of expertise in writing. In Rijlaarsdam, G., Alamargot, D. & Chanquoy, L. (Vol. Eds.), *Studies in writing: Vol. 9. Through the models of writing*. Dordrecht, The Netherlands: Kluwer Academic Publishers. 219-228.

Kellogg, R. T. & Raulerson, B. A. (2007). Improving the writing skills of college students. *Psychonomic Bulletin & Review, 14*(2), 237–242. https://doi.org/10.3758/BF03194058.

Kellogg, R.T., Turner, C., Whiteford, A. & Mertens, A. (2016). The role of working memory in planning and generating written sentences." *The Journal of Writing Research 7.*(3) 397-416.

Khezrlou. (2021). Effects of task repetition with written corrective feedback on the knowledge and written accuracy of learners with different prior knowledge of the structure. *Revista Española de Lingüística Aplicada*, *34*(2), 464–493. https://doi.org/10.1075/resla.19054.khe

Khezrlou. (2021). Focus on form in task repetition through oral and written task modeling. *International Review of Applied Linguistics in Language Teaching,* IRAL, 61(2), 479–518. https://doi.org/10.1515/iral-2020-0125

Khezrlou, S. (2019). Task repetition and corrective feedback: The role of feedback types

and structure saliency. *English Teaching and Learning*, *43*(2), 213–233. https://doi.org/10.1007/s42321-019-00025-2

Khezrlou, S. (2020). The role of task repetition with direct written corrective feedback in L2 writing complexity, accuracy and fluency. *Journal of Second Language Studies*, *3*(1), 31–54. https://doi.org/10.1075/jsls.19025.khe

Kim, A.-Y. A., & Kim, H. J. (2017). The effectiveness of instructor feedback for learning-oriented language assessment: Using an integrated reading-to-write task for English for academic purposes. Assessing Writing, 32, 57-71. https://doi.org/10.1016/j.asw.2016.12.001

Kim, J. & Li, S. (2024). The effects of task repetition and corrective feedback on L2 writing development. *The Language Learning Journal*, 1–16. https://doi.org/10.1080/09571736.2024.2390555

Kim, Y. & Tracy-Ventura, N. (2013). The role of task repetition in L2 performance development: What needs to be repeated during task-based interaction? *System (Linköping)*, *41*(3), 829–840. https://doi.org/10.1016/j.system.2013.08.005

Kim, Y., Choi, B., Yun, H., Kim, B., & Choi, S. (2022). Task repetition, synchronous written corrective feedback and the learning of Korean grammar: A classroom-based study. *Language Teaching Research*, *26*(6), 1106–1132. https://doi.org/10.1177/1362168820912354

Kim, Y. J., Crossley, S., Jung, Y. J., Kyle, K., & Kang, S. (2018). The effects of task repetition and task complexity on L2 lexicon use. In *Task-Based Language Teaching* (Vol. 11, pp. 75-96). John Benjamins Publishing Company. https://doi.org/10.1075/tblt.11.03kim

Knowles, M. (1970. *The modern practice of adult education, (Volume 41).* New York: New York Association Press.

Kormos, J. (2014). Differences across modalities of performance: An investigation of linguistic and discourse complexity in narrative tasks. In H. Byrnes & R.M. Manchon (Eds.), *Task- based language learning insights from and for L2 writing,* 193-216. Amsterdam: John Benjamins Publishing Company.

Kormos, J. (2006). *Speech production and L2 acquisition.* Mahwah, NJ: Lawrence Erlbaum.

Kormos, J. (2011). Task complexity and linguistic and discourse features of narrative writing performance. *Journal of Second Language Writing*, *20*(2), 148–161. https://doi.org/10.1016/j.jslw.2011.02.001

Kormos, J., Brunfaut, T. & Michel, M. (2020) Motivational factors in computer-administered integrated skills tasks: A study of young learners, *Language Assessment Quarterly, 17*(1), 43-59, DOI: 10.1080/15434303.2019.1664551

Kruk, M., & Zawodniak, J. (2018). Boredom in practical English language classes: Insight

from interview data. In L. Szyma´nski, J. Zawodniak, A. Łobodziec, & M. Smoluk (Eds.), *Interdisciplinary views on the English language, literature and culture, 177–191. Uniwersytet Zielonogórsk.*

Kuiken, F., & Vedder, I. (2008). Cognitive task complexity and written output in Italian And French as a foreign language. *Journal of Second Language Writing, 17,* 48–60. doi:10.1016/j.jslw.2007.08.003

Kuiken, F., & Vedder, I. (2012). Syntactic complexity, lexical variation and accuracy as a function of task complexity and proficiency level in L2 writing and speaking. In Housen, A., Kuiken, F. & Vedder, I. (eds.), *Dimensions of L2 Performance and Proficiency. Complexity, Accuracy* and Fluency in SLA. 143-170. John Benjamins.

Kuiken, F., Vedder, I., Gilabert, R. (2010). Communicative adequacy and linguistic complexity in L2 writing. In Barting, I., Martin, M., Vedder, I. (Eds.), Communicative proficiency and linguistic development: Intersections between SLA and language testing research. EuroSLA Monographs 1. (pp. 81–100). Rome: EuroSLA.

Kyle, K. (2021). Measuring longitudinal writing development using indices of syntactic complexity and sophistication. *Studies in second language acquisition, 43,* 781 – 812.

Kyle. (2020). The relationship between features of source text use and integrated writing quality. *Assessing Writing, 45*, 100-467. https://doi.org/10.1016/j.asw.2020.100467

Kyle, K. (2023, March 19). *The reliability and validity of lexical diversity indices in second language oral productions.* [Conference presentation]. AAAL 2023 Conference, Portland, Or., United States. https://www.aaal.org/events/aaal-2023-conference---portland-oregon

Kyle, K. & Eguchi, M. (2023). Assessing spoken lexical and lexicogrammatical proficiency using features of word, bigram, and dependency bigram use. The Modern Language Journal, 107, 531-564. https://doi.org/10.1111/modl.12845

Kyle, K., Crossley, S. A., & Jarvis, S. (2021). Assessing the validity of lexical diversity using direct judgements. *Language Assessment Quarterly 18*(2), pp. 154-170. https://doi.org/10.1080/15434303.2020.1844205

Kyle, K., Sung, H., Eguchi, M., & Zenker, F. (2024). Evaluating evidence for the reliability and validity of lexical diversity indices in L2 oral task responses. *Studies in Second Language Acquisition*, *46*(1), 278–299. doi:10.1017/S0272263123000402

Lado, R. (1957). *Linguistics across cultures: Applied linguistics for language teachers*. University of Michigan Press.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology, 4,* 863. https://doi.org/10.3389/fpsyg.2013.00863

Lambert, C. & Oliver, R. (2020). *Using tasks in second language teaching. Practice in diverse contexts.* Cambridge University Press.

Lambert, C., Kormos, J., & Minn, D. (2016). Task repetition and second language speech processing. *Studies in Second Language Acquisition, 39*(1), 167–196. https://doi.org/10.1017/S0272263116000085

Lambert, C., Philp, J., & Nakamura, S. (2017). Learner-generated content and engagement in second language task performance. *Language Teaching Research : LTR*, *21*(6), 665–680. https://doi.org/10.1177/1362168816683559

Lan, G. et al. (2019). Grammatical complexity: 'what does it mean' and 'so what' for L2 writing classrooms? *J. Second Lang. Writ.* 46:100673. doi: 10.1016/j.jslw.2019.100673

Larsen-Freeman, D. (2009). Adjusting expectations: The study of complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics, 30*, 579–589.

Larsen-Freeman, D. (1983). Assessing global second language proficiency. In Seliger, H. & Long, M. (Eds), *Classroom-oriented research in second language acquisition.* Newbury House. 287-304.

Larsen-Freeman, D. (2000). Techniques and principles in language teaching, 2nd edition. Oxford University Press.

Larsen-Freeman, D. (2006). The Emergence of Complexity, Fluency, and Accuracy in the Oral and Written Production of Five Chinese Learners of English. *Applied Linguistics, 27*, 590-619.

Laufer, B. & Nation, P. (1995). Vocabulary Size and Use: Lexical Richness in L2 Written Production. *Applied Linguistics, 16*(3), 307-322. Lee, J. (2018. Task complexity, cognitive load, and L1 speech. Applied Linguistics, 1–35. https://doi. org/10.1093/applin/amx054.

Levelt, W. J. (1999). Producing spoken language: A blueprint of the speaker. In C. Brown & P. Hagoort (Eds.), Neurocognition of language, 83–122. Oxford: Oxford University Press.

Levelt, W. J., & Speaking, M. (1989). *From intention to articulation*. The MIT Press.

Lewkowicz, J. A. (1997). *The integrated testing of a second language.* In C. Clapham & D. Corson (Eds.), Encyclopedia of language and education: Language testing and assessment 7 (pp. 121-130). Dordrecht, the Netherlands: Kluwer Academic.

Levy, C. M., & Ransdell, S. E. (1995). Is writing as difficult as it seems? *Memory and Cognition, 23*, 767–779

Li, P., & Rogers, J. (2021, December 6). Task repetition in language learning: A bibliography. *OSF Preprints. https://doi.org/10.31219/osf.io/rk7ag*

Lightbown, P. M. (2019). Perfecting practice. *Modern Language Journal, 103*(3), 703–712. https://doi.org/10.1111/modl.12588

Lightbown, P. & Spada. N., (1993). How Languages Are Learned, Oxford: Oxford University Press.

Lin, O. & Maarof, N. (2013). Collaborative writing in summary writing: Student perceptions and problems. Procedia-Social and Behavioral Sciences, 90, 599 – 606.

Lindqvist, C., Gudmundson, A., & Bardel, C. (2013). A new approach to measuring lexical sophistication in L2 oral production. *EuroSLA.* 109 – 126.

Lintunen, P. & Makila, M. (2014). Measuring syntactic complexity in spoken and written learner language: Comparing the incomparable? *Research in Language, 12*(4). 377 – 399.

Loewen, S. (2015). Introduction to instructed second language acquisition. New York: Routledge.

Long, M. (1985). A role for instruction in second language acquisition: Task-based language teaching. In K. Hyltenstam and M. Pienemann (Eds.), Modelling and assessing second language acquisition (pp. 77-99). Clevedon, Avon: Multilingual Matters.

Long, M. (2015). Second language acquisition and task-based language teaching. Malden, MA: Wiley Blackwell.

Long, M., & Norris, J. (2000). Task-Based teaching and assessment. In M. Byram (Ed.), *Encyclopedia of language teaching* (pp. 597–603). Routledge.

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writer's language development. TESOL Quarterly, 45, 36–61.

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal, 96*(2), 190–208. https://doi.org/10.1111/j.1540-4781.2011.01232.x

Lund, R.J. (1990). A taxonomy for teaching second language listening. *Foreign Language Annals,* 23, 105-115.

Lynch, T. & Maclean, J. (2000). Exploring the benefits of task repetition and recycling for classroom language learning. *Language Teaching Research, 4*, 221-250.

Maamuujav, U. (2021). Examining lexical features and academic vocabulary use in adolescent L2 students' text-based analytical essays. *Assessing Writing*, *49*, 100540-. https://doi.org/10.1016/j.asw.2021.100540

Mackey, A. & Gass, S. (2005). *Second language research: methodology and design.* Lawrence Erlbaum.

Mady, C. & Seiling, A. (2018). Functional adequacy distinguishes immigrant multilinguals in French speaking task. *International journal of arts, humanities and social sciences, 3*(4). 1 – 7.

Martínez, A. C. L. (2018). Analysis of syntactic complexity in secondary education ELF writers at different proficiency levels. *Assessing Writing, 35*, 1-11.

Maru, G., Pikirang, C., & Liando, N. Integrating writing with listening in EFL class: A systematic review. *Advances in social science, education and humanities research, 473,* 222-226.

Master, P. (1997) The English article system: Acquisition, function, and pedagogy. *System, 25* (2), 215-232.

McCarthy P. & Jarvis S. (2010). A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods, 42*(2), 381–392. [PubMed]

Manchón, R.M. & Matsuda, P. (2018). *Handbook of second and foreign language writing.* Berlin, Germany: Walter de Gruyter.

Manchón, R.M. (2014). The distinctive nature of task repetition in writing. Implications for theory, research, and pedagogy. *Estudios de Lingüística Inglesa Aplicada (ELIA), 14,* 13-41.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of Sophisticated approaches to lexical diversity assessment. *Behavior Research Methods, 42,* 381–392.

McCulloch, S. (2013). Investigating the reading-to-write processes and source use of L2 postgraduate students in real-life academic tasks: An exploratory study. *Journal of English for Academic Purposes, 12*(2), 136-147.

McHugh, M. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*(3), 276-282.

McNamara, T. F. (2000). *Language testing*. Oxford: Oxford University Press.

Mekala, S., Ponmani, M., & Shabitha, M. (2016). Transfer of grammatical knowledge into ESL Writing. Eurasian Journal of Applied Linguistics, 2(2), 47 – 64.

Mennim, P. 2003. Rehearsed oral L2 output and reactive focus on form. *ELT Journal, 57*(2), 130–38. doi:10.1093/elt/57.2.130.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan Publishing Co, Inc; American Council on Education.

Michel, M. (2017). Complexity, accuracy and fluency in L2 production. In S. Loewen, & M. Sato (Eds.), *Routledge Handbook of Instructed Second Language Acquisition.* Routledge. 50 – 68.

Michigan Language Assessment. (2020). *What is the CEFR Framework?* https://michiganassessment.org/wp-content/uploads/2020/02/20.02.PDF.CEFR-Framework-Basics.pdf

Miller, J. (2005). Most of ESL students have trouble with the articles. *International Education Journal, 5*, 80-88.

Mislevy, R., Steinberg, L., & Almond, R. (2002). Design and analysis in task-based language assessment. *Language Testing*, *19*(4), 477–496. https://doi.org/10.1191/0265532202lt241oa

Muhammadpour, M., Hassanzadeh-Taleshi, M., & Salehi-Amiri, F. (2023). The effects of different task repetition schedules on oral narratives of L2 learners with high and low working memory capacity. *Acta Psychologica*, *236*, 103933–103933. https://doi.org/10.1016/j.actpsy.2023.103933

Murakami, A., & Alexopoulou, T. (2016). L1 influence on the acquisition order of English grammatical morphemes: A learner corpus study. *Second Language Research, 32*(3), 365–394. https://doi.org/10.1177/0267658315621049

Nambiar, R. (2007). Enhancing academic literacy among tertiary learners: a Malaysian experience. 3L Journal of Language Teaching, Linguistics and Literature, 13.

Narmina, R. (2024). Error correction as part of a teaching process. Bulletin of the Bohdan Khmelnytskyi National University of Cherkasy. Series_"Pedagogical Sciences", 2, *100-106.* https://doi.org/10.31651/2524-2660-2024-2-100-106

Neumann, H. (2014). Teacher assessment of grammatical ability in second language academic writing: A case study. *Journal of Second Language Writing, 24,* 83–107.

Newmeyer F. J., Preston L. B. (Eds). (2014). *Measuring Grammatical Complexity.* Oxford: Oxford University Press; 10.1093/acprof:oso/9780199685301.001.0001

Nguyen, H., Nguyen, H., Vo, N. & Huynh, N. (2023). The combined effects of task repetition and post-task teacher-corrected transcribing on complexity, accuracy and fluency of L2 oral performance. *International Review of Applied Linguistics in Language Teaching*. 1-29. https://doi.org/10.1515/iral-2023-0128

Nitta, R. & Baba, K. (2018). Understanding benefits of repetition from a complex dynamic systems perspective. The case of a writing task. In Martin Bygate (ed.), *Learning language through task repetition*, 285–316. Amsterdam: John Benjamins.10.1075/tblt.11.11nit

Noroozi, M. & Taheri, S. (2022) Task-based language assessment: a compatible approach to assess the efficacy of task-based language teaching vs. present, practice, produce, *Cogent Education*, 1-21. DOI: 10.1080/2331186X.2022.2105775

Norouzian, S., Yaqubi, B., & Ahmadpour Kasgari, Z. (2023). Task repetition from EFL learners' perspectives: A longitudinal multiple-case analysis. *Language Related Research, 14*(3), 123–144. https://doi.org/10.29252/LRR.14.3.5

Norris, J. M. (2016). Current uses for task-based language assessment. *Annual Review of Applied Linguistics*, *36*, 230–244. https://doi.org/10.1017/ S0267190516000027

Norris, J. M. (2002). Interpretations, intended uses and designs in task-based language assessment. *Language Testing*, *19*(4), 337–346. https://doi.org/10.1191/ 0265532202lt234ed

Norris, J. M. (2018). Task-based language assessment aligning designs with intended uses and consequences. *JLTA Journal*, *21*, 3–20 doi:https://doi. org/10.20622/jltajournal.21.0_3.

Norris, J. M., Brown, J. D., Hudson, T. D., & Bonk, W. (2002). Examinee abilities and task difficulty in task-based second language performance assessment. *Language Testing*, *19*(4), 395–418. https://doi.org/10.1191/0265532202lt237oa

Norris, J. M., & East, M. (2021). Task-based language assessment. In M. J. Ahmadian & L. M. H. (Eds.), *The Cambridge handbook of task-based language teaching* (pp. 507–528). Cambridge University Press.

Norris, J.M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics, 30*, 555–578.

Nunan, D. (1997). Designing and Adapting Materials to Encourage Learner Autonomy. In P. Benson, & P. Voller (Eds.). *Autonomy and Independence in Language Learning,* 192-203. London: Longman.

Nunan, D. (1989). Designing Tasks for the Communicative Classroom. Cambridge: Cambridge University Press.

Odlin, T. (1989). *Language transfer: Cross-linguistic influence in language learning*. Cambridge University Press.

Ohta, R., Plakans, L., & Gebril, A. (2018). Integrated writing scores based on holistic and multi-trait scales: A generalizability analysis. Assessing Writing, 38, 21-36.

Ong, J., & Zhang, L. J. (2010). Effects of task complexity on the fluency and lexical complexity in EFL students' argumentative writing. *Journal of Second Language Writing, 19,* 218-233

Ormrod, J. E. (2011). *Educational psychology: developing learners.* 7th ed. Boston, Pearson/Allyn & Bacon.

Ortega, L. (1999). Planning and focus on form in L2 oral performance. *Studies in Second Language Acquisition, 21*, 109 - 148.

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.

Othman, N. (2009). Teaching and assessing three types of direct writing in Malaysian ESL classrooms – a survey of ESL teachers' opinions. *English Language Journal,* 3, 102 – 114.

Pallant, J. (2020). *SPSS Survival Manual: A step by Step-by-Step Guide to Data Analysis Using IBM SPSS*. McGraw-Hill Education.

Pallotti, G. (2015). A simplistic view of linguistic complexity. *Second Language Research, 31*(1), 117-134.

Pallotti, G. (2019). Assessing tasks: The case of interactional difficulty. *Applied Linguistics, 40*(1), 176-197.

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics, 30*(4), 590–601.

Patanasorn. (2010). *Effects of procedural content and task repetition on accuracy and fluency in an EFL context*. ProQuest Dissertations Publishing.

Payant, C., Mcdonough, K., Uludag, P., & Lindberg, R. (2019). Predicting integrated writing task performance: Source comprehension, prewriting planning, and individual differences. *Journal of English for Academic Purposes*, *40*, 87–97. https://doi.org/10.1016/j.jeap.2019.06.001

Plakans, L. (2010). Independent versus integrated writing tasks: A comparison of task representation. *Tesol Quarterly, 44,* 185-194.

Plakans, L. (2009). The role of reading strategies in integrated L2 writing tasks. Journal of English for Academic Purposes, 8(2), 252 – 266.

Plakans, L., & Gebril, A. (2012). Discourse complexity in integrated writing tasks. *Assessing Writing, 17*(3), 148–166. https://doi.org/10.1016/j.asw.2012.05.001

Plakans, L., & Gebril, A. (2013). Using multiple sources in a listening and reading-based writing assessment task: Source text use as a predictor of score. *Second Language Writing Journal, 22,* 217–230.

Plakans, L., Gebril, A., & Bilki, Z. (2019). Shaping a score: Complexity, accuracy, and fluency in integrated writing performances. *Language Testing, 36*(2), 161 – 179.

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning, 64*(4), 878–912. https://doi.org/10.1111/lang.12079

Plough, I., & Gass, S. (1993). Interlocutor and task familiality: Effects of interactional structure. In G. Crookes & S. Gass (Eds.), *Tasks and language learning. 35–56).* Multilingual Matters.

Polio, C. (1997). *Measures of linguistic accuracy in second language writing research. Language Learning, 47,* 101–143.

Polio, C. (2001). Research methodology in second language writing researching: The case of text-based studies. In T. J. Silva & P. K. Matsuda (Eds.), *On second language writing*, *91–115.* Mahwah, NJ: Lawrence Erlbaum.

Prabhu, N.S. (1987). *Second language pedagogy. Oxford: Oxford University Press.*

Qin, J. & Liu, D. (2024) Introducing/Testing New SFL-Inspired Communication/Content/Function-Focused Measures for Assessing L2 Narrative Task Performance, *Applied Linguistics*, https://doi.org/10.1093/applin/amae030

Qin, X. Q., Wen Q. F. (2007). *EFL Writing of College English Majors in China: A*

Developmental Perspective. Beijing: China Social Sciences Press.

Ransdell, S., Arecco, M., & Levy, C. (2001). Bilingual long-term working memory: The effects of working memory loads on writing quality and fluency. *Applied Psycholinguistics, 22*(1), 113-128.

Read, J. (1990). Providing relevant content in an EAP writing test. *English for Specific Purposes, 9*(2), 109–121.

Révész, A. (2011). Task Complexity, Focus on L2 constructions, and individual differences: A classroom-based study. *Modern Language Journal, 95*(Supplement s1), 162-181.

Révész, A, Kourtali, N-E and Mazgutova, D (2017) Effects of Task Complexity on L2 Writing Behaviors and Linguistic Complexity. *Language Learning, 67*(1). pp. 208-241.

Richard, J., Platt, J., & Weber, H. (1985). *Longman dictionary of applied linguistics*. London: Longman.

Rinehart, S & Thomas, K. (1993). Summarization ability and text recall by novice studiers. *Reading Research and Instruction, 32*(4). 24 – 32.

Robbins, R. (2011). Assessment and accountability of academic advising. In Joslin, J. & Markee, N. (Eds.), *Academic advising administration: Essential knowledge and skills for the 21st century (Monograph No. 22)*. National Academic Advising Association. 53–64.

Robinson, P. (2005). Cognitive complexity and task sequencing: Studies in a componential framework for second language task design. *International Review of Applied Linguistics, 43,* 1 – 32.

Robinson, P. (Ed.). (2011). Second language task complexity: Researching the Cognition Hypothesis of language learning and performance. Amsterdam: John Benjamins. Robinson, P. (2001a). Task complexity, cognitive re-sources, and syllabus design: A triadic frame-work for investigating task influences on SLA. In P. Robinson (Ed.), *Cognition and second language instruction.* 287–318. New York: Cambridge University Press.

Robinson, P. & Gilabert, R. (2007). Task complexity, the cognition, hypothesis and second language learning and performance. *Language Testing. 45*(2), 161 – 176.

Rokoszewska, K. J. (2022). The dynamics of monthly growth rates in the emergence of complexity, accuracy, and fluency in L2 English writing at secondary school–a learner corpus analysis. *System, 106*, Article 102775.

Roothooft, H., Lázaro-Ibarrola, A., & Bulté, B. (2022). Task repetition and corrective feedback via models and direct corrections among young EFL writers: Draft quality and task motivation. *Language Teaching Research : LTR*, *29*(3), 1311–1344. https://doi.org/10.1177/13621688221082041

Rost, M. (2002) *Teaching and Researching Listening.* Harlow: Pearson Education Limited.

Rukthong, A. (2016). *Investigating the listening construct underlying listening-to-summarize tasks*. [Doctoral Thesis, Lancaster University]. Lancaster University.

Rukthong, A. & Brunfaut, T. (2020). Is anybody listening? The nature of second language listening in integrated listening-to-summarize tasks. *Language Testing*. 1 – 23.

Sakai, H. (2009). Effect of Repetition of Exposure and Proficiency Level in L2 Listening Tests. *TESOL Quarterly*, *43*(2), 360–372. http://www.jstor.org/stable/27785016

Sakuragi, T. (2011). The construct validity of the measures of complexity, accuracy, and fluency: Analyzing the speaking performance of learners of Japanese. *JALT Journal, 33,* 157–173.

Sample, E & Michel, M. (2014). An exploratory study into trade-off effects of complexity, accuracy, and fluency on young learners' oral task repetition. TESL Canada Journal, 31(8), 23 – 46.

Samuda, V., & Bygate, M. (2008). *Tasks in second language learning*. Palgrave Macmillan. https://doi.org/ 10.1057/9780230596429

Sang, Z., & Zou, W. (2023). The Effect of Joint Production on the Accuracy and Complexity of Second Language Writing. *Journal of psycholinguistic research*, *52*(2), 425–443. https://doi.org/10.1007/s10936-022-09882-8

Segalowitz, N. 2010. The cognitive bases of second language fluency. New York: Routledge.

Shin, S., & Ewert, D. (2015). What accounts for integrated reading-to-write task scores? Language Testing, 32(2), 259-281.

Shintani, N., & Ellis, R. (2013). The comparative effect of direct written corrective feedback and metalinguistic explanation on learners' explicit and implicit knowledge of the English indefinite article. *Journal of Second Language Writing, 22*, 286–306.

Shintani, N., Ellis, R., & Suzuki, W. (2014). Effects of written feedback and revision on learners' accuracy in using two English grammatical structures. *Language Learning, 64,* 103–131.

Skehan, P. (1998a). *A cognitive approach to language learning.* Oxford: Oxford University.

Skehan, P. (1996). A framework for the implementation of task-based instruction, *Applied Linguistics*, 17(1). 38–62.

Skehan, P. (1989). *Individual Indifferences in Second Language Learning*. London: Edward Arnold.

Skehan, P. (2009). Modelling L2 performance: Integrating complexity, accuracy, fluency and lexis. *Applied Linguistics, 30(4),* 510–532.

Skehan, P. (Ed.) (2014). *Processing perspectives on task performance.* Amsterdam: John Benjamins.

Skehan, P. (1998b). Task-based instruction. *Annual Review of Applied Linguistics*, *18*. 268-286.

Skehan, P. & Foster. P. (2001). Cognition and tasks. In Robinson, R. (Ed.), *Cognition and second language instruction*. 183.205. New York: Cambridge University Press.

Skehan, P. & Foster. P (1997). Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research, 1*(3), 185-231.

Specia, L., Jauhar, S.K., & Mihalcea, R. (2012). Semeval-2012 task 1: English lexical simplification. In The 1st *SEM.

Storch, N. & Wigglesworth, G. (2010). Learners' processing, uptake and retention of corrective feedback on writing: Case studies. *Studies in Second Language Acquisition 32*(2), 303–334.

Storch, N. & Wigglesworth, G. (2006). Writing tasks: the effects of collaboration. In M.D.P. Mayo (Ed.) *Investigating tasks in formal language learning.* 157-177. Clevedon: Multilingual Matters Ltd.

Suzuki, M. (2017) Complexity, accuracy, and fluency measures in oral pre-task planning: A synthesis. *Second Language Studies 36*(1), 1–52.

Tabari, M., Khezrlou, S., & Tian, Y. (2022). Task complexity, task repetition, and L2 writing complexity: exploring interactions in the TBLT domain. *International Review of Applied Linguistics in Language Teaching, IRAL*. https://doi.org/10.1515/iral-2022-0123

Tabari, M., Khezrlou, S., & Tian, Y. (2023). Verb argument construction complexity indices and L2 written production: effects of task complexity and task repetition. *Innovation in Language Learning and Teaching*, *ahead-of-print*(ahead-of-print), 1–16. https://doi.org/10.1080/17501229.2023.2211955

Tavakoli, P. (2009). Investigating task difficulty: learners' and teachers' perceptions. *International Journal of Applied Linguistics, 19*(1), 1-25.

Tavakoli, P., & Hunter, A.-M. (2017). Is fluency being 'neglected' in the classroom? Teacher understanding of fluency and related classroom practices. Language Teaching Research, 22(3), 330–349. https://doi.org/10.1177/1362168817708462

Tavakoli, P. (2014). Storyline complexity and syntactic complexity in writing and speaking tasks. In H. Byrnes & R.M. Manchon (Eds.), *Task-based language learning insights from and for L2 writing,* 217-236. Amsterdam: John Benjamins Publishing Company.

Taylor, C., & Angelis, P. (2008). The evolution of the TOEFL. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.). *Building a validity argument for the TOEFL iBT,* 27–54. New York, NY: Taylor and Francis.

Teng, M. F., & Huang, J. (2021). The effects of incorporating metacognitive strategies instruction into collaborative writing on writing complexity, accuracy, and fluency. *Asia Pacific Journal of Education*, *43*(4), 1071–1090. https://doi.org/10.1080/02188791.2021.1982675

Thai, C., & Boers, F. (2016). Repeating a monologue under increasing time pressure: Effects on fluency, complexity, and accuracy. *TESOL Quarterly, 50*(2), 369–393. https://doi.org/10.1002/tesq.232

Thirakunkovit, S., & Rhee, S. (2021). Grammatical Complexity as a Predictor of Difficulty of Grammar Items in an English Test. *THAITESOL JOURNAL*, *34*(2), 930118. https://doi.org/10.1186/s40468-021-00144-3

Van den Branden, K., Bygate, M., & Norris, J. M. (Eds.) (2009). *Task-based language teaching: A reader*. Amsterdam: John Benjamins.

Van den Branden, K. (Ed.). (2006). *Task-based language education: From theory to practice*. Cambridge University Press.

VanPatten, B. (2012). *Input processing.* In S. M. Gass & A. Mackey (Eds.), The handbook of second language acquisition. 268-281). New York: Routledge.

Vermeer, A. (2004). The relation between lexical richness and vocabulary size in Dutch L1 andL2 children. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a Second Language: Selection, Acquisition and Testing.* (pp. 173-189). (Language Learning and Language Teaching; No. 10). John Benjamins Publishing Company.

Wang, X. & Jin, C. (2022). Effects of task complexity on linguistic complexity for sustainable EFL writing skills development. *Sustainability, 14.* 1 – 14.

Watanabe, Y. (2001). *Read-to-write tasks for the assessment of second language academic writing skills: Investigating text features and rater reactions.* Unpublished doctoral dissertation, University of Hawaii.

Way, D. P., Joiner, E. G., & Seaman, M. A. (2000). Writing in the secondary foreign Language classroom: The effects of prompts and tasks on novice learners of French. *Modern Language Journal, 84,* 171–184.

Weigle, S. C. (2004). Integrating reading and writing in a competency test for non-native speakers of English. *Assessing Writing, 9*(1), 27–55.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach.* Antony Rowe Ttd.

Widdowson, H. (1998). The Theory and Practice of Critical Discourse Analysis. *Applied Linguistics, 19,* 136-151.

Willis, J. (1996). *A framework for task-based learning*. Harlow: Longman.

Willis, D., & Willis, J. (2007). *Doing Task-Based Teaching.* Oxford: Oxford University Press.

Willis, D. & Willis, J. (2001). Task- based language learning. In R. Carter & D. Nunan (Eds.), *The Cambridge to teaching ro speakers of other languages* (pp. 173-79). Cambridge: Cambridge University Press.

Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). *Second language development in writing: Measures of fluency, accuracy and complexity.* Manoa: University of Hawaii Press.

Yang, H. and Plakans, L. (2012), Second language writers' strategy use and performance on an integrated reading-listening-writing task. TESOL Quarterly, 46: 80–103.

Yang, Y., & Lyster, R. (2010). Effects of form-focused practice and feedback on Chinese EFL learners' acquisition of regular and irregular past tense forms. *Studies in Second Language Acquisition, 32,* 235–263.

Yang, Y., & Zheng, Z. (2024). A refined and concise model of indices for quantitatively measuring lexical richness of Chinese university students' EFL writing. *Contemporary Educational Technology, 16*(3), ep513. https://doi.org/10.30935/cedtech/14707

Zabihi, R. (2018). The role of cognitive and affective factors in measures of L2 writing. *Written Communication, 35*(1), 32-57.

Zenker, F., & Kyle, K. (2021). Investigating minimum text lengths for lexical diversity indices. *Assessing Writing, 47*, https://doi.org/10.1016/j.asw.2020.100505

Zhang, M., Yi, N., & Zhou, D. (2023). The Effects of Task Repetition Schedules on L2 Fluency Enhancement. *Languages (Basel)*, *8*(4), 252-. https://doi.org/10.3390/languages8040252

Zhang, S., Zhang, H., & Zhang, C. (2022). A Dynamic Systems Study on Complexity, Accuracy, and Fluency in English Writing Development by Chinese University Students. *Frontiers in psychology*, *13*, 787710. https://doi.org/10.3389/fpsyg.2022.787710

Zhao, Z., Nimehchisalem, V., & Chan, M. Y. (2024). Independent Writing Tasks vs. Integrated Writing Tasks: The Cognitive Demands and Their Impacts on Linguistic Complexity in EFL Writing. *Theory and Practice in Language Studies*, *14*(8), 2606–2617. https://doi.org/10.17507/tpls.1408.33

Zhu, X., Li, X., Yu, G., Cheong, C. M., & Liao, X. (2016). Exploring the Relationships Between Independent Listening and Listening-Reading-Writing Tasks in Chinese Language Testing: Toward a Better Understanding of the Construct Underlying Integrated Writing Tasks. *Language Assessment Quarterly*, *13*(3), 167–185. https://doi.org/10.1080/15434303.2016.1210609

Zúñiga, C. (2016). Implementing task-based language teaching to integrate language skills in an EFL program at a Colombian university. *PROFILE Issues in Teachers' Professional Development, 18*(2), 13-27. http://dx.doi.org/10.15446/profile.v18n2.49754.

Appendices

Appendix 1: Participant information sheet

English second language speakers' performances on listening-to-write tasks

Participant information sheet for students

My name is John Bandman, and I am a PhD student at Lancaster University, U.K. I would like to invite you to take part in my PhD study: "English second language speakers' performances on listening-to-write tasks."

Please take time to read the following information carefully before you decide whether or not you wish to take part.

What is the study about?
This study aims to explore English second language speakers' performances of listening-to-write tasks.

Why have I been invited?
I have approached you because you are in an English as a Second Language course, and I am interested in understanding how students complete listening-to-write tasks. I would be very grateful if you would agree to take part in this study.

Do I have to take part?
No. It's completely up to you to decide whether or not you take part. Your participation is voluntary. If you decide not to take part in this study, this will not affect your studies and the way you are assessed in your classes.

Will my data be identifiable?
After I collect the data, only I, the researcher conducting this study and my supervisor from Lancaster University will have access to the data. I will keep all personal information about you (e.g. your name and other information about you that can identify you) confidential, that is I will not share it with others. I will encrypt all names, and I will remove all personal information. In publications and presentations, I will use pseudonyms if I have to refer to an individual person

What will I be asked to do if I take part?
If you decided to take part, you will first be asked to complete a personal background questionnaire that takes about 10 minutes to complete, providing information on, for example, your age, first language background, and English language learning experience. I will then ask you to complete three English listening-to-write tasks that take 20 minutes each over a 6-week time period, and afterwards a short questionnaire on your experience of doing the tasks, which takes about 10 minutes to complete. You may also be asked to take part in a short interview on your experiences of completing the tasks.

What are the possible benefits from taking part?
Completing this project will help English as a Second Language teachers improve the ways they teach writing. In turn, it helps teachers become better at helping present and future English second language learners improve their writing. You may also enjoy doing these tasks, which can be seen as extra language practice.

What if I change my mind?

If you change your mind, you have the right to withdraw from the project at any time, but no later than thirty (30) days after you have given consent to take part in the study. If you decide not to take part in this study, this will not affect your studies and the way you are assessed in your classes. However, beyond thirty days, it is impossible to take out data from one specific participant when this has already been encrypted and pooled together with other people's data.

What are the possible disadvantages and risks of taking part?
It is unlikely that there will be any major disadvantages to taking part.

How will my data be stored?
Your data will be stored in encrypted files (that is no-one other than me, the researcher and my supervisor will be able to access them) and on password-protected computers.
I will store hard copies of any data securely in locked cabinets in my work office.
I will keep data that can identify you separately from non-personal information (e.g. your views on a specific topic). In accordance with Lancaster University guidelines, I will keep the data securely for a minimum of ten years.

How will you use the information you have shared with me and what will happen to the results of the research study?
I will use the data you have shared with me only in the following ways: I will use it for academic and professional purposes only. This will include my thesis and potentially academic and professional journal/book publications. I may also present the results of my study at academic and professional conferences, and use examples from the data in my teaching. If you give me permission to look up your placement test score, I will be able to verify your placement level and make possible correlations when reviewing the results from the study.

When writing up the findings from this study, I will mainly report the results at the general level, and I may discuss some of the views and ideas you shared with me. When doing so, I may use your exact words, but I will use pseudonyms to refer to you.

Who has reviewed the project?
This study has been reviewed and approved by the Faculty of Arts and Social Sciences and Lancaster Management School's Research Ethics Committee.

What if I have a question or concern?
If you have any queries or if you are unhappy with anything that happens concerning your participation in the study, please contact myself, John Bandman at Bergen Community College, Ender Hall E-122, 400 Paramus Road, Paramus, N.J. 07652, United States of America, Tel +19176121411 **j.bandman@lancaster.ac.uk** or my supervisor, Dr. Tineke Brunfaut at Lancaster University, Department of Linguistics and English Language, County South, LA1 4YL, Lancaster, United Kingdom, Tel: +44(0)1524 594084, **t.brunfaut@lancaster.ac.uk**
If you have any concerns or complaints that you wish to discuss with a person who is not directly involved in the research, you can also contact the Head of Department: Professor Elena Semino, Lancaster University, Department of Linguistics and English Language, County South, LA1 4YL, Lancaster, United Kingdom, Tel: +44(0)1524 594176, e.semino@lancaster.ac.uk

Thank you for considering your participation in this project.

Appendix 2: Consent form

Project Title: English second language speakers' performances on listening-to-write tasks.

Name of Researcher: John Bandman
Email: j.bandman@lancaster.ac.uk

Please tick each box

1. I confirm that I have read and understand the information sheet for the above study. I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily ☐
2. I understand that my participation is voluntary. I also understand the information on withdrawing as described in the information sheet. ☐
3. If I am participating in the classroom, I understand that any information disclosed within the classroom remains confidential to the group, and I will not discuss the study with or in front of anyone who was not involved unless I have the relevant person's express permission. ☐
4. Since this study will take place in a classroom setting, I understand that any information given by me may be used in future reports, academic articles, publications, presentations or teaching by the researcher, but my personal information will not be included and I will not be identifiable. ☐
5. I understand that my name will not appear in any reports, articles or presentation without my consent. ☐
6. I understand that any interviews will be audio-recorded and transcribed and that data will be protected on encrypted devices and kept secure. ☐
7. I understand that data will be kept according to University guidelines for a minimum of 10 years after the end of the study. ☐
8. I am giving you permission to look up my placement test score. ☐
9. I agree to take part in the above study. ☐

Name of Participant	Date	Signature
_____	_____	_____

I confirm that the participant was given an opportunity to ask questions about the study, and all the questions asked by the participant have been answered correctly and to the best of my ability. I confirm that the individual has not been coerced into giving consent, and the consent has been given freely and voluntarily.

Signature of Researcher /person taking the consent_____

Date_____ Day/month/year

One copy of this form will be given to the participant and the original kept in the files of the researcher at Lancaster

Appendix 3: Pre-task questionnaire: Student demographics

Note: Students completed the Qualtrics online version through this following link: https://eu.qualtrics.com/jfe/form/SV_8wsqcUzl0oZwdBb

Q1 Please write your first name.

_____

Q2 Please write your last name.

_____

Q3 Please write your e-mail address.

_____

Q4 What is your gender?

- ○ Male
- ○ Female
- ○ Other

Q5 What is your age (in years)?

_____

Q6 What is your first/native language?

_____

Q7 What is your nationality?

_____

Q8 How many years have you been studying English?

_____

Q9 How many years have you lived in English-speaking or English-dominant countries?

_____

Q10 How would you rate your English language skills?

	Beginner	Basic	Intermediate	Upper-Intermediate	Advanced
Speaking	○	○	○	○	○
Listening	○	○	○	○	○
Reading	○	○	○	○	○
Writing	○	○	○	○	○
Vocabulary	○	○	○	○	○
Grammar	○	○	○	○	○

Q11 Which of the following describes your current student status?

- ○ Undergraduate
- ○ Graduate

Q12 If you are an undergraduate student, which year of study are you in at your university?

- ○ 1st
- ○ 2nd
- ○ 3rd
- ○ 4th or more

Q13 If you are a graduate student, which degree are you currently pursuing?

- ○ Masters
- ○ PhD

Q14 What is your current area of study?

- ○ Humanities (Communication, History, Religion, Philosophy, English, ESL & World Languages)
- ○ Art (Fine Arts, Performing Arts, Visual Arts, Architecture, Fashion, Photography)
- ○ Business (Marketing, Accounting, Entrepreneurship, Hotel/Restaurant Management, Business)
- ○ Legal (Criminal Justice, Law or Security)
- ○ Social Sciences (Psychology, Sociology, Human Development, Social Work)
- ○ Health Professions (Nursing, Pre-Med, Medicine)
- ○ Mathematics
- ○ Education
- ○ Technology
- ○ Environmental Sciences
- ○ Physical Sciences

Other or Undeclared

Appendix 4: Listening input transcript

http://www.esl-lab.com/class/classc1.htm

Teacher: Okay, Okay, let's begin. Hello, everyone. My name's Karl Roberts, and I'll be your teacher for this class, Intercultural Communication 311.

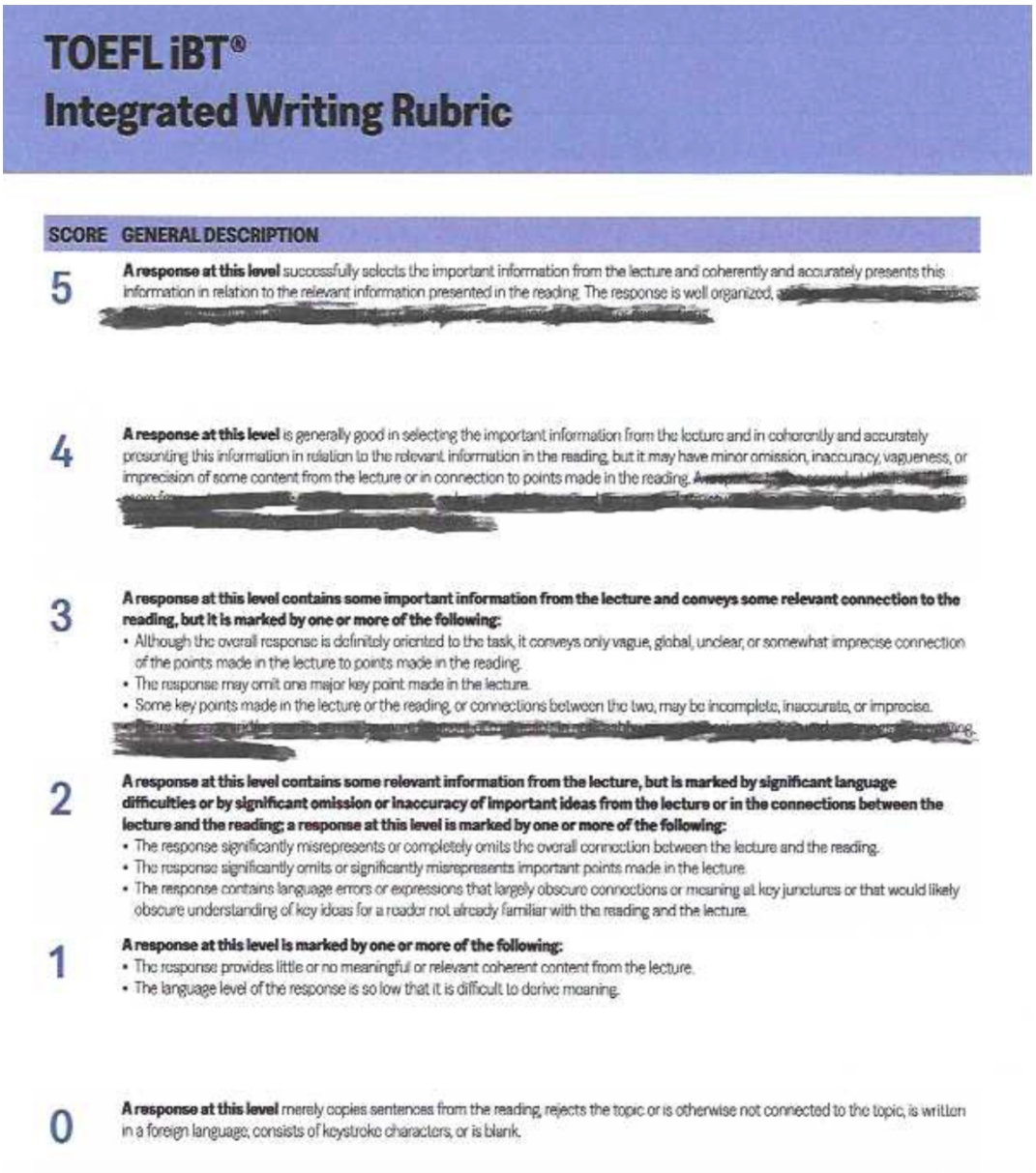
To begin with, uh, please look at the syllabus in front of you. You should all have one by now, I think. This class meets on Tuesdays and Thursdays from 3:15 to 4:50. We will be meeting in this room for the first half of the course, but we will be using the research lab every other week on Thursday in room 405 during the last two months of the class.

Uh, this is the text for the class, Beyond Language. Unfortunately, the books haven't come in yet, but I was told that you should be able buy them at the bookstore the day after tomorrow. Again, as you see on your course outline, grading is determined by your work on a midterm and final test, periodic quizzes, uh, a research project, and classroom participation.

My office hours are from 1:00 to 2:00 on Wednesdays, and you can set up an appointment to meet with me at other times as well. Okay, let me explain a little bit more about the class and its objectives.

Appendix 5: Scoring sheet: Integrated writing rubric (knowledge summary)

[Downloaded from https://www.ets.org/pdfs/toefl/toefl-ibt-writing-rubrics.pdf]



TOEFL iBT®
Integrated Writing Rubric

SCORE	GENERAL DESCRIPTION
5	**A response at this level** successfully selects the important information from the lecture and coherently and accurately presents this information in relation to the relevant information presented in the reading. The response is well organized, [redacted]
4	**A response at this level** is generally good in selecting the important information from the lecture and in coherently and accurately presenting this information in relation to the relevant information in the reading, but it may have minor omission, inaccuracy, vagueness, or imprecision of some content from the lecture or in connection to points made in the reading. [redacted]
3	**A response at this level contains some important information from the lecture and conveys some relevant connection to the reading, but it is marked by one or more of the following:** • Although the overall response is definitely oriented to the task, it conveys only vague, global, unclear, or somewhat imprecise connection of the points made in the lecture to points made in the reading. • The response may omit one major key point made in the lecture. • Some key points made in the lecture or the reading, or connections between the two, may be incomplete, inaccurate, or imprecise. [redacted]
2	**A response at this level contains some relevant information from the lecture, but is marked by significant language difficulties or by significant omission or inaccuracy of important ideas from the lecture or in the connections between the lecture and the reading; a response at this level is marked by one or more of the following:** • The response significantly misrepresents or completely omits the overall connection between the lecture and the reading. • The response significantly omits or significantly misrepresents important points made in the lecture. • The response contains language errors or expressions that largely obscure connections or meaning at key junctures or that would likely obscure understanding of key ideas for a reader not already familiar with the reading and the lecture.
1	**A response at this level is marked by one or more of the following:** • The response provides little or no meaningful or relevant coherent content from the lecture. • The language level of the response is so low that it is difficult to derive meaning.
0	**A response at this level** merely copies sentences from the reading, rejects the topic or is otherwise not connected to the topic, is written in a foreign language, consists of keystroke characters, or is blank.

Appendix 6: Rating scale for knowledge transfer

(Adapted from Association of American Colleges and Universities)

Notes:

1. I used only the yellow highlighted criterion in my study.
2. The numbers represent the score band descriptors.

4

3

2

1

INTERDISCIPLINARY LEARNING VALUE RUBRIC

This rubric was developed by the AAC&U and is used here as a sample of how the learning objectives might be assessed. For more information on the development of the Essential Skills rubrics, please contact info@aacu.org

While departments and programs should NOT alter learning objectives, *performance indicators* used to measure the learning objectives may be altered and can vary depending upon the course or program. **Departments will need to indicate the performance indicators used to measure each of the 2-3 learning objective (per essential skill) assessed in their program reviews.**

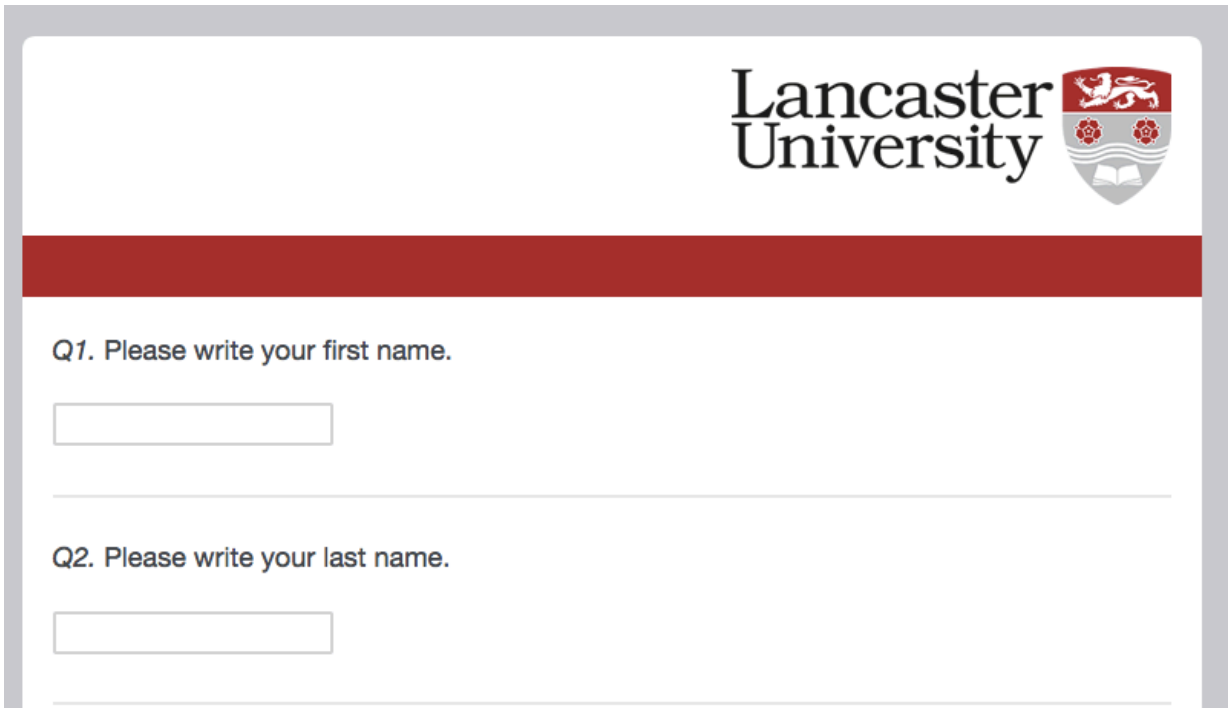
Definition

Interdisciplinary learning is an understanding and a disposition that a student builds across the curriculum and cocurriculum, from making simple connections among ideas and experiences to synthesizing and transferring learning to new, complex situations within and beyond the campus.

Learning Objectives	Most advanced performance indicators			Least advanced performance indicators
Connections to Experience *Connects relevant experience and academic knowledge*	Meaningfully synthesizes connections among experiences outside of the formal classroom (including life experiences and academic experiences such as internships and travel abroad) to deepen understanding of fields of study and to broaden own points of view.	Effectively selects and develops examples of life experiences, drawn from a variety of contexts (e.g., family life, artistic participation, civic involvement, work experience), to illuminate concepts/theories/frameworks of fields of study.	Compares life experiences and academic knowledge to infer differences, as well as similarities, and acknowledge perspective other than own.	Identifies connections between life experiences and those academic texts and ideas perceived as similar and related to own interests.
Connections to Discipline *Sees (makes) connections across disciplines, perspectives*	Independently creates wholes out of multiple parts (synthesizes) or draws conclusions by combining examples, facts, or theories from more than one field of study or perspective.	Independently connects examples, facts, or theories from more than one field of study or perspective.	When prompted, connects examples, facts, or theories from more than one field of study or perspective.	When prompted, presents examples, facts, or theories from more than one field of study or perspective.
Transfer *Adapts and applies skills, abilities, theories, or methodologies gained in one situation to new situations*	Adapts and applies, independently, skills, abilities, theories, or methodologies gained in one situation to new situations to solve difficult problems or explore complex issues in original ways.	Adapts and applies skills, abilities, theories, or methodologies gained in one situation to new situations to solve problems or explore issues.	Uses skills, abilities, theories, or methodologies gained in one situation in a new situation to contribute to understanding of problems or issues.	Uses, in a basic way, skills, abilities, theories, or methodologies gained in one situation in a new situation.
Integrated Communication	Fulfills the assignment(s) by choosing a format, language, or graph (or other visual representation) in ways that enhance meaning, making clear the interdependence of language and meaning, thought, and expression.	Fulfills the assignment(s) by choosing a format, language, or graph (or other visual representation) to explicitly connect content and form, demonstrating awareness of purpose and audience.	Fulfills the assignment(s) by choosing a format, language, or graph (or other visual representation) that connects in a basic way what is being communicated (content) with how it is said (form).	Fulfills the assignment(s) (i.e to produce an essay, a poster, a video, a PowerPoint presentation, etc) in an appropriate form.
Reflection and Self-Assessment *Demonstrates a developing sense of self as a learner, building on prior experience to respond to new and challenging contexts (may be evident in self-assessment, reflective, or creative work)*	Envisions a future self (and possibly makes plans that build on past experiences that have occurred across multiple and diverse contexts).	Evaluates changes in own learning over time, recognizing complex contextual factors (e.g., works with ambiguity and risk, deals with frustration, considers ethical frameworks).	Articulates strengths and challenges (within specific performances or events) to increase effectiveness in different contexts (through increased self-awareness).	Describes own performances with general descriptors of success and failure.

Appendix 7: Task repetition perception questionnaire

https://eu.qualtrics.com/jfe/form/SV_bvo6ITuYOH7jGN7



Q3. Questions about this task. For each of the following statements, please choose the column that best represents your level of agreement.

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
I enjoyed repeating the listening-to-write task.	○	○	○	○	○
Repeating the listening-to-write task was boring.	○	○	○	○	○
The task reflects my English writing ability.	○	○	○	○	○
The task accurately reflects my English listening ability.	○	○	○	○	○
I had enough time to complete the writing task.	○	○	○	○	○
The audio recording was played enough times for me to understand it.	○	○	○	○	○
The audio recording provided sufficient ideas for me to complete the writing task.	○	○	○	○	○
The audio recording was too long.	○	○	○	○	○
Vocabulary in the audio recording was difficult for me.	○	○	○	○	○
Sentence structures in the audio recording were complicated for me.	○	○	○	○	○

Q4. Questions about tasks in general. For each of the following statements, please choose the column that best represents your level of agreement.

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
Repeating a task helps me improve my writing.	○	○	○	○	○
Repeating a task helps me improve my listening.	○	○	○	○	○
Writing after listening improves my writing.	○	○	○	○	○
Listening with the purpose of writing helps me improve my English.	○	○	○	○	○
Fulfilling a listening-to-write task gets easier with repetition.	○	○	○	○	○
I would like to do task repetition in future classes.	○	○	○	○	○

Q5. What is your overall opinion about task repetition? Explain.

Appendix 8: Error-code correction symbols

(provided by the participating university)

Correction Symbols

Code	Meaning	Example	Correction
S/V	**Subject-Verb Agreement** *Subject and verb don't agree.*	They has a new car; Tin drive it.	They have a new car; Tin drives it.
VF	**Verb Form**	I am work everyday.	I work everyday.
VT	**Verb Tense**	She was studying everyday.	She studies everyday.
GI	**Gerund Infinitive** *Gerund or infinitive mistake.*	I need going home. I look forward to see you.	I need to go home. I look forward to seeing you.
SP	**Spelling**	I like this bok.	I like this book.
# Sing/pl	**Singular/Plural**	I saw a lot of dog. Every dogs was brown.	I saw a lot of dogs. Every dog was brown.
WP	**Wrong Pronoun**	Them are very nice.	They are very nice.
WO	**Word Order**	It's a problem economic.	It's an economic problem.
WF	**Word Form** – *Wrong form – adj, adv, noun, verb.*	The kindness man helped her.	The kind man helped her.
MW	**Missing Word**	I like candy because is sweet.	I like candy because it is sweet.
WW	**Wrong word**	I assisted the concert.	I attended the concert.
PREP	**Preposition** – *missing/mistake*	I went the doctor's office.	I went to the doctor's office.
art	**Article** – *missing/mistake* (a, an, the)	She bought new sweater.	She bought a new sweater.
poss	**possessive**	My dad sister is here.	My dad's sister is here.
C	**Capitalization**	my name is raoul.	My name is Raoul.
Frag/inc	**fragment/incomplete**	He asked where my sister.	He asked where my sister was.