GENERATION AND L2 LEARNING: AN INVESTIGATION INTO GENERATION AS A DESIRABLE DIFFICULTY IN L2 LEARNING CONTEXTS AMONG UNIVERSITY LEARNERS IN JAPAN

by

JOHN ANTHONY DUPLICE

A thesis submitted to the University of Birmingham for the degree of

DOCTOR OF PHILOSOPHY

Department of English Language and Applied Linguistics

College of Arts and Law

University of Birmingham

November 2023

ACKNOWLEDGEMENT

Looking back over the years since I started this PhD endeavor, I realise that I could not have completed this thesis, nor even have come close, without the help and guidance of several individuals. I am profoundly grateful to my supervisors, Dr. Gareth Carrol, Dr. Petra Schoofs, and Dr. Paul Thompson, for their continuous support and guidance throughout the course of my research. Their deep expertise, insightful feedback, and patient mentorship have been invaluable to my work.

In particular, I would like to extend my gratitude to my primary advisor, Dr. Gareth Carrol, without whom I would never have finished or started this PhD thesis. Dr. Carrol`s willingness to respond and provide support to my questions, many of which I asked multiple times, was invaluable throughout the years of my research. I was regularly amazed at his willingness and ability to provide detailed feedback on drafts or explain my mistakes in coding, of which there were many, so quickly and with great patience. After receiving his feedback, whether it was through email or during a Zoom meeting, I felt a burst of often-needed excitement and confidence in my ability to complete my research and thesis. From the day I first met Dr. Carrol at a conference in Japan to today as I submit my thesis, he has been the foundation of the experience and a true mentor.

I would also like to express my heartfelt gratitude to my writing research group Caroline Hadley and Tom Gorham and the Japanese Association of Language Teachers Mind, Brain, and Education SIG. The intellectually stimulating discussions and camaraderie have made this journey a truly enriching experience. Their perspectives and suggestions have greatly enriched my research. I would also like to thank Tracey Tokuhama-Espinosa for rejuvenating

my passion for teaching and challenging me through her class on the neuroscience of learning, which put me on a path toward the research in this present thesis.

My deepest thanks go to my family for their support and patience throughout this endeavor. This journey of pursuing a PhD has been both challenging and rewarding. It has been a period of intense learning, not just about my field of study, but also about perseverance and resilience.

ABSTRACT

This thesis investigated the efficacy of Generation used as a desirable difficulty (DD) in a second language (L2) context. This was a modular thesis that consisted of three modules, the present module being the third and final. Building upon an initial study conducted in Module Two, this final module describes three studies conducted among university English language learners (ELLs) in Japan. This thesis examined the impact of Generation on L2 learning, focusing on vocabulary, reading comprehension, and self-feedback in academic writing.

The first study looked at the efficacy of Generation on L2 vocabulary meaning recall over three and 15-week periods. This study investigated the impact of Generation tasks on long-term learning among 38 university ELLs in Japan. The study compared the Generation of novel sentences using target L2 vocabulary to a baseline of recall tasks. This was done to investigate whether the benefits found in L1 contexts would carry over to L2 vocabulary learning. While the study did show some interactions that show a trend to possible benefits of Generation, the study did not find a significant clear benefit to using Generation in the L2 context among the participants studied.

Study 2 investigated the application of Generation tasks in L2 reading comprehension, examining both L1 and L2 Generation usage through prompts embedded in the reading text. The study consisted of 44 university ELLs in Japan and explored the impact of varying cognitive load demands through two distinct Generation tasks and a control condition. This novel approach aimed to understand the influence of Generation on L2 reading comprehension and the comparative effects of employing Generation tasks in L1 versus L2. The results did not

indicate a benefit from Generation in either L1 or L2. Moreover, L2 Generation incurred a cost compared to the control condition of rereading the text.

Study 3 explored the role of self-reflection in L2 academic writing among 39 ELL university students, utilizing Generation checklists. The study analyzed three aspects of writing - Formatting, Grammar, Spelling, Punctuation, and Capitalization (GSPC), and Content - across four essays. The results indicated that Generation led to improvements in all three writing aspects compared to the control condition, suggesting that Generation tasks through self-reflection can enhance L2 academic writing. The study also examined the carryover effect of Generation tasks from one essay to the next. Content scores significantly improved in the final two essays, indicating that the deeper thinking invoked by Generation tasks in earlier essays may have enhanced subsequent learning and ability.

The findings from these present studies revealed that Generation's efficacy in L2 contexts differs from previous results found in L1 contexts and can vary within L2 contexts and tasks. While Generation was found ineffective as an initial DD in vocabulary learning and for L2 reading comprehension, it showed benefits when used as self-feedback in L2 writing.

Table of Contents

List of Tables

List of Figures

Terminology and Typographical Conventions

Action Research

- Action research, as defined in this thesis, empowers educators to investigate and improve their teaching practices, fostering a cycle of reflection and action leading to better student learning for a specific setting that may or may not be extended to other educational contexts.

ELL (English Language Learner):

- An ELL is an individual learning English as a second or additional language. ELLs may have various levels of proficiency in English, and they often require specialized language instruction to improve their English language skills.

ESL (English as a Second Language):

- ESL refers to the programs, courses, or instructional approaches designed to teach English to individuals whose first language is not English. ESL programs are typically offered in countries where English is the dominant language.

EFL (English as a Foreign Language):

- EFL refers to the teaching and learning of English in a non-English-speaking country. In EFL contexts, English is learned as a subject rather than as a means of communication in daily life.

L1 (First Language):

- L1, also known as the mother tongue, is the language that a person learns as their first language, typically from birth or early childhood.

L2 (Second Language):

- L2 refers to a language that is learned after the first language (L1). It can be any language other than the individual's first language.

SLA (Second Language Acquisition):

- SLA is the process through which individuals learn and acquire a second or additional language, such as English. It encompasses the development of language skills, including speaking, listening, reading, and writing, in a language that is not L1.

Implicit and Explicit Learning:

- Implicit Learning: This is a subconscious or incidental form of learning in which individuals acquire knowledge or skills without conscious awareness. In language learning, it often involves picking up grammar and vocabulary through exposure and usage.

- Explicit Learning: This is a conscious and deliberate form of learning where individuals actively study and acquire knowledge or skills through instruction and focused practice. In language learning, it may involve explicit grammar rules and vocabulary drills.

Learner-Centered / Teacher-Centered:

- Learner-centered: An instructional approach that places the learner at the center of the learning process, emphasizing their needs, interests, and active engagement in learning. The teacher serves as a facilitator.

- Teacher-centered: An instructional approach that places the teacher in a central role, with the teacher directing and controlling the learning process. Learners often follow a prescribed curriculum and receive direct instruction.

Word Families:

- Word families consist of groups of words that share a common root or base word and have similar meanings or functions. As in this thesis, they are often used in vocabulary instruction to help learners understand and use related words.

Rote Memorization:

- Rote memorization involves learning information through repetition and memorization without necessarily understanding the underlying concepts. It is often characterized by the memorization of facts or information without deeper comprehension.

Fluency Effect / False Fluency:

- Fluency effect / false fluency: This refers to the phenomenon where learners may perceive themselves as more fluent in a language or skill than they actually are. It can result from overconfidence or a lack of awareness of one's limitations. This also occurs when a learner can produce language or perform a skill fluently in controlled practice situations but struggles to do so in more natural, real-life contexts.

Metacognition:

- Metacognition is the ability to think about and control one's own cognitive processes. It involves self-awareness and self-regulation of one's thinking, learning, and problem-solving strategies. This is often seen during self-reflective activities.

Schemas:

- Schemas are mental frameworks or organized structures that individuals use to organize and make sense of information. In language learning, learners develop schemas for understanding and using language rules and patterns.

Scaffolding:

- Scaffolding is a teaching technique where instructors provide support and guidance to learners as they work on tasks or concepts, which is often reduced as students gain proficiency.

CHAPTER 1: INTRODUCTION

1.0 Introduction

Acquiring a second language (L2) takes a great deal of effort by the learner to communicate effectively. While there is no easy way to avoid difficulty learning a second language, the methods a learner implements play an important role in how long it takes and to what level they succeed in the target L2. However, there are a variety of methods and approaches to learning a second language in an L2 university classroom, where student variables and classroom dynamics make no two learning environments the same. This third module of a modular PhD aims to identify and evaluate a set of specific approaches and strategies known as desirable difficulties (DDs) in acquiring a second language by investigating L2 university learners in Japan. This chapter will introduce the investigation by initially discussing the background and context, followed by the research problem, the research aims, the significance, and the organization of the thesis.

1.1 The Disconnect Between Research and the Classroom

In Japan, English language classes are compulsory from grade five in public elementary schools (M.E.X.T., 2020), and many private schools start English education earlier. English education continues through secondary school and is a key part of tests for high school and university entrance exams. However, Japan consistently ranks low in English ability worldwide, and its 2022 English Proficiency Index (EF EPI, 2022) ranking was 80th, down from 78th in 2021, despite English being a core subject in compulsory education. Moreover, upon entering university, students often continue to struggle to communicate in English and

often find themselves relearning content covered in schooling prior to entering university. One reason for this is the focus on external motivating factors around how English ability is assessed. In the lead-up to the Japanese university EFL context, many students have studied English to pass a test rather than to learn for long-term language use (Otomo & Danping, 2016). These tests look to assess vocabulary knowledge, reading comprehension, and short-term listening skills through multiple-choice question tests. To perform well on these tests, students study, and teachers teach to the test, leading to a short-term test-taking ability that often does not transfer to communicating effectively in a broader context. While there is a need to change the external motivation of test focus through compulsory education, there is also an immediate need for effective study techniques that will support long-term retention. Students must also be familiar with effective methods, and teachers need to implement the methods in the curriculum. Methods identified within the field of learning sciences can provide the techniques for teachers to implement and students to use in the ELL classroom.

The field of learning sciences, also known as the science of learning, is interdisciplinary in nature as it encompasses the processes of the study of learning throughout educational settings. It further concentrates on learning research in practical settings, such as the classroom. While research into how people learn is well established in fields such as educational psychology, learning sciences look beyond the general theories of learning and explore specific concepts in practical settings like the classroom, where measuring the efficacy of the concept is often more difficult than in a laboratory setting. Tokuhama-Espinosa (2019) describes learning sciences as a field that explains the most effective learning methods under different conditions and how certain variables influence learning outcomes. This need is evident in the classroom, where factors stemming from the students, instructor, or classroom settings can be

innumerable; therefore, having a better understanding of how certain factors affect learning outcomes when using different methods is of great benefit to both instructors and learners.

While the efficacy of different learning methods is often well understood by researchers in fields such as learning sciences, psychology, and neuroscience, this understanding does not consistently make its way to the classroom where it is most needed. This is seen in the focus on entrance exams of English and how they influence the methods that students use, and the way that teachers focus on teaching to pass a test over long-term learning for communication (Ryan, 2008). Therefore, there is a need for effective learning methods for ELLs in Japanese universities, since these students will likely have used methods for short-term goals through tertiary EFL education. One such set of methods is the addition of difficulty to tasks to enhance learner attention and long-term retention. These methods are known as desirable difficulties. This thesis will investigate the use of desirable difficulties in the EFL university classroom in Japan and their role in effective L2 learning.

1.2 What is Known

Methods of study are an aspect of education that can be affected when there is a disconnect between research findings and the classroom. This is seen in the encoding of new material or content, a key part of the learning process. Encoding, as defined in this thesis, builds upon Shunk`s (2012) definition of constructing and storing knowledge in memory, but expands upon it to the taking of knowledge collected in short-term or working memory, and moving it to long-term memory where it can be recalled hours, days, or months later as needed. An example commonly seen in the classroom is when reading text for learning, students often use rereading and highlighting as encoding methods (Agarawal & Bain, 2019). However, it has

long been understood in the research literature that these are inefficient in encoding content (Brown, Roediger, & McDaniel, 2014; Weinstein, Sumeracki, & Caviglioli, 2018). There are a few reasons why rereading and highlighting alone are not efficient. For highlighting, learners are focused on marking facts or points they think are important, but this step alone does not make connections or help better understand concepts. The problem with rereading is that it provides more exposure to the target material but not more depth (Bjork & Bjork, 2011). This leads to a familiarity with the material but not necessarily an understanding of it, or an ability to transfer it to different materials later. This familiarity can lead the learner to a false belief that they know the target material when, in fact, they only have a very superficial familiarity with it. Callender and McDaniel (2009) showed little advantage in short-term retention and no benefit in longer-term recall of reading content among university students in an L1 context. This study was particularly important in that it looked at both memory and learning through different types of academic texts, such as textbooks and scientific articles. If research has shown highlighting and rereading to be ineffective, then why do students continue to use these methods when studying? Didau (2015) suggests that one reason for this and other inefficient study methods that students use is that they think what feels most comfortable is the most effective. This is further strengthened when tested soon after studying, as familiarity can help in correctly answering test questions when shallow understanding is all that is needed (Otomo & Danping, 2016; Soderstrom & Bjork, 2015). This is where desirable difficulties might offer a beneficial alternative.

1.3 Defining Desirable Difficulties

Desirable difficulties (DDs) enhance learning skill acquisition by adding a certain amount of difficulty to a learning task. This added difficulty comes at a cost that often diminishes immediate or short-term results. In this chapter, the difference between short-term and long-term results is discussed in verbal learning. DDs are implemented in acquiring two broad types of skills, verbal and motor skills. Verbal skills are concerned primarily with cognition, such as language learning, reading, and mathematics, while motor skills often relate to muscle movements, such as shooting baskets in basketball, and dance movements. This thesis investigates DDs and verbal learning and, therefore, will not address the use of motor skills beyond periodic examples to compare with verbal learning. Henceforth, unless stated otherwise, any discussion of skills and DDs refers to verbal skills.

1.3.1 Performance Versus Learning

Desirable difficulties make learning more effortful, but the difficulty leads to improved longer-term results. These positive results are what make the difficulty desirable. Simply making something harder when studying does not equate to a desirable difficulty. According to Bjork and Bjork (2011), learners must possess the appropriate background knowledge or skills to incorporate the difficulty to make it desirable, otherwise it will become an undesirable difficulty and may hinder learning. DDs are based on the distinction between two concepts, *performance,* and *learning*. Soderstrom and Bjork (2015) describe performance as the short-term observable outcome during or immediately after a studying activity. Learning, on the other hand, is a relatively permanent change in knowledge. An example of this difference can be seen in the classroom. Students will learn about and initially grasp a new concept taught by

an instructor. Immediately upon completing the task, the instructor will give them a quiz on the material to verify that students understood. This is considered performance, as the material is still fresh in the students` minds soon after exposure. While the learners may have been able to perform on the initial post-exposure quiz, they may very well not be able to remember the target content in a post-test a few weeks later, and therefore, the content would not have been considered learned (Soderstrom & Bjork, 2015). In the L2 context of English language learning, the use of prepositions "in, at, on" for locations in a city may be taught to students in a lecture format. Soon after the instruction on the use of these prepositions, students will likely be able to use them correctly in an immediate post-lecture quiz, which would be an example of performance. However, if the use of the prepositions is tested a few weeks later, the ability may no longer be present, illustrating a lack of learning. Conversely, as learning is more permanent, it can lead to the long-term retention of content and the ability to transfer the learned material to new learning or different contexts. If learning takes place, students can correctly use the target prepositions weeks after being introduced and could transfer the preposition used to new learning.

A key concept to understanding the distinction between performance and learning is *blocked* (or massed) studying versus distributed learning. Blocked time sessions entail studying a large amount of target content in a single block of time, a practice commonly referred to as "cramming." This is often observed in university settings before a test, where students aim to memorize as much content as possible within a limited timeframe, such as the night before the test. Following this blocked session, students must show what they know on a test or other assessment. Although students may demonstrate adequate performance on a test by recollecting the facts they studied during the block session, their retention of the material is typically short-

lived (see Cepeda et al., 2009; Soderstrom & Bjork, 2015). Consequently, both the instructor and learner may lack insight into what is retained weeks following the test, thus hindering future recall and application in subsequent learning activities.

While learners may be able to perform well by remembering many of the facts they studied in the prior blocked session, they will likely forget the content soon afterward, as blocked material is often fleeting (Soderstrom & Bjork, 2015). This may account for the issues that Japanese ELLs face as the study techniques they use are often focused on performance on tests over longer-term learning. If the study sessions were distributed over a longer period of time leading up to the initial quiz, the material studied would likely be more durable and, therefore, more likely to be recalled in a future post-test and used in future learning (see Brown et al., 2014; Karpicke & Bauernschmidt, 2011; Nakata, 2019). In learning English, Japanese students in junior and senior high school often study to the test. This can be a result of the *backwash (washback) effect*, a phenomenon in educational testing that refers to the influence of testing on curriculum design, teaching practices, and learning behaviors. This brings up the consequential validity of the effects of the specific tests on how instruction by the teacher and studying by the student is undertaken and the social implications that develop as a result (Allen & Nagatomo, 2019; Brown, 1997; Sudjasmada, 2021). As high-stakes English entrance tests primarily focus on surface-level grammar understanding, vocabulary recognition, and basic listening comprehension, the education is focused on performing in these areas, none of which require generating content in English. Thus, performance through blocked studying is adequate to pass the tests, and long-term learning is often less of a focus.

The alternative to blocked studying is distributed learning, which involves spacing out study sessions over a more extended period leading up to a test. This has been shown to

facilitate better encoding and retention of information, enhancing the learner`s ability to recall and apply the material in future learning tasks (Brown, Roediger, & McDaniel, 2014; Soderstrom & Bjork, 2015). While distributed learning can play a role in longer-term retention, it will not affect what is being learned; therefore, the backwash effect of studying to the test would likely continue. What could be alleviated is the need to relearn the same material repeatedly, as it would be more likely to have been moved to long-term memory.

Blocked practice, encompassing blocked time sessions and blocked content, involves studying a single material type in isolation (Richland et al., 2005). While it may enhance short-term performance, its effects are often transient (Soderstrom & Bjork, 2015). In contrast, interleaved learning involves studying mixed content types, enhancing long-term recall and application skills due to the induced DD (Bird, 2011; Dunlosky et al., 2013; Rohrer & Taylor, 2007). Nakata and Suzuki (2019) demonstrated that while blocked practice yielded better immediate post-quiz results in L2 grammar studies, interleaved practice proved superior in long-term assessments. Cepeda et al. (2008) corroborated these findings in both L1 and L2 contexts. Their research indicated that without regular review sessions, the advantages of blocked practice are short-lived compared to distributed practice, where benefits on recall are more enduring.

Another issue arising from performance is *false fluency* (also known as fluency illusion or the fluency effect). When the learner performs well on the initial quiz soon after the blocked study session, they and the instructor may have a false belief in their ability or knowledge of the studied material. This false understanding in the learning will cause future difficulty when it is necessary to transfer previously learned material to new contexts. In L2 learning, students may study a prepared set of words or dialog that can be performed in class, leading to false

confidence in the ability to use the target language outside of the classroom setting. Among university students, less difficult learning methods (e.g., blocked) can lead to false confidence in knowledge, while more challenging learning methods (e.g., interleaved) instils a more accurate level of confidence (Agarawal & Bain, 2019; Bjork & Bjork, 1992; Brown et al., 2014). Willingham (2021) sums up the learner`s difference in perspective between performance and learning, where the brain encourages the learner to do what is comfortable (e.g., blocked practice), giving the impression of progress toward success. However, in reality, more difficult mental exercises lead to more benefits in the long run.

The difference between performance and learning is that performance reflects short-term temporary gains, and learning is more durable in long-term knowledge acquisition or skills. Furthermore, blocked and interleaved practices were introduced in how they are reflected in performance and learning. This thesis will explore performance and learning as these concepts relate to individual DDs from different contexts. Specifically, a DD known as generation is investigated as its role and efficacy are not clear in L2 learning.

1.4 Generation Introduction

Generation (also known as the generation effect) requires learners to engage deeply with content. This results in a slower pace of study as it is particularly challenging. Generation necessitates not only active engagement on the part of the learner but also creativity in generating meaning, examples, or content related to the material being studied. The use of elaboration in learning is a form of generation as the learner describes or explains aspects of the material being generated. This is often done by making connections between the content being learned and previous experiences or knowledge (Weinstein et al., 2018). Another perspective

on generation is that it involves the learner making sense of new information by relating or testing it against their existing knowledge (Enser & Enser, 2020). Learning science research has demonstrated that knowledge generation through active engagement with content (e.g., textbooks, word lists, graded readers) can improve recall ability in various L1 contexts and settings (Bertsch et al., 2007; McCurdy et al., 2020). In contrast, less active strategies such as highlighting and rereading are considered to have low utility and limited benefit (Dunlosky, Rawson, Marsh, Nathan & Willingham, 2013). The difference in benefit is often attributed to increased difficulty stemming from a rise in cognitive processes from implementing generation in the learning process (Bjork, 2011; Roediger & Karpicke, 2006). Generation has been studied extensively in L1 contexts and has been found to show a positive effect in different learning situations while being less effective in others. What is not evident in the literature is the effect of generation on L2 learning. Therefore, this thesis looks to address this need by exploring generation on L2 learning from multiple aspects.

1.5 Thesis Aims and Overview

The primary aim of this research is to investigate the efficacy of generation in the L2 context. This will be achieved by looking at generation from three different areas of focus: vocabulary recall, reading comprehension, and self-feedback from reflection in writing. The significance of this research lies in its potential to contribute to the relatively limited body of knowledge on the use of generation among L2 learners in the university classroom. To achieve these aims, this thesis employs three empirical action research studies (as well as discussing a prior action research study completed as an earlier part of this PhD research) to investigate three aspects of L2 learning by using generation tasks. The action research studies were

conducted among students in the classrooms of the instructor / researcher, which led to limitations in the studies. The limitations included small sample sizes, participants that were predetermined by the university through English-level placement tests, constraints limiting access to participants, inability to balance gender, and extrinsic factors (e.g., high-stakes tests in other classes and student club activities) that may have affected student attention. Additionally, this thesis was limited to a population of Japanese university students at a specific university in Tokyo, Japan, which may affect the generalizability of the findings.

The first area of research investigates L2 vocabulary learning, an area in the literature with findings ranging from mixed to negative results. Additionally, studies conducted to date have primarily focused on shorter-term recall. This thesis looks at vocabulary from a longer-term perspective, providing valuable insights to compare with previous findings that looked at the shorter term. One prior study discussed in Chapter 3 and one novel study in Chapter 4 are employed to investigate the benefits of generation on vocabulary learning in more detail.

The second area of research investigates L2 reading comprehension using generative tasks. The literature shows some effects of generation in different L1 reading situations. However, this is less clear in the L2 context. Additionally, the research conducted in this study used a novel approach of using L1 as a condition for reducing cognitive load, which is discussed in Section 2.4.1, to determine the effect of generation on reading tasks.

The final study is an exploratory investigation into generation of self-reflection in L2 academic writing through post-writing checklists. Although limited research has been conducted on this specific topic, there are parallels with other studies on reflective writing, such as those examining self-assessment and peer feedback in L2 writing contexts. These studies provide a foundation for further exploration and comparison.

1.6 Organization of the Thesis

This thesis consists of seven chapters that are centered around four experimental studies. This chapter has presented the topic and provided background information on the disconnect between research and the classroom. It also introduced desirable difficulties and set up the problem of performance versus learning. The chapter defined desirable difficulties and introduced the concept of generation, the primary DD investigated in this thesis. The aims and research questions of the thesis and their significance are also provided at the end of this chapter. Chapter 2 provides a general literature review on second language acquisition, L2 vocabulary learning, and desirable difficulties such as spacing, interleaving, retrieval, feedback, disfluency, and generation. Counterarguments to DDs from Cognitive Load Theory are also discussed. Chapter 3 provides an overview of Modules 1 and 2 and reviews the initial vocabulary study conducted in Module 2, setting the stage for another study on L2 vocabulary meaning recall discussed in Chapter 4. This chapter presents the findings of the second vocabulary study, referred to as Study 1 in this thesis. Chapter 5 consists of a study on generation in L2 reading, looking at a novel approach to investigating cognitive load on reading comprehension using generation. Chapter 6 presents the results and implications of a more exploratory writing checklist study incorporating generation. The final chapter, Chapter 7, summarizes and discusses the findings of the studies, leading to recommendations for future research.

2.0 CHAPTER 2: LITERATURE REVIEW

2.1 Introduction to Desirable Difficulties in SLA

DDs encompass multiple strategies to improve learning by adding difficulty to the learning process (Bjork, 1994). It is important to note that the keyword here is desirable, since not all difficulties are desirable or effective in enhancing learning. The difficulty must be carefully designed and implemented to be effective in long-term learning, and, therefore, desirable (Persellin & Daniels, 2018; Weinstein, Sumeracki, & Caviglioli, 2018). If the difficulty is too difficult, it may interfere with the learning process or cause a lack of motivation among the learners.

The slowing down of the learning process is a key aspect of the effectiveness of DDs, as this can lead to deeper cognitive processing of the target content, leading to improved longer-term learning (Bjork, 1994). However, DDs are often less effective for short-term recall and can require the learner to understand the benefits before implementing DD strategies (Soderstrom & Bjork, 2015). This is because self-reported belief in the effectiveness of DDs shows a disconnect in that learners often believe that immediate improvement (performance) is evidence of learning (Roberts & Kreuz, 2015). Roberts and Kreuz provide other examples of a disconnect in L2 learning methods for oral communication with speakers of the target language, such as using flashcards, drill practice, and recorded lessons. This can lead to a phenomenon known as negative transfer, where the way of learning interferes with real-world language use (2015). The lack of accurate student judgment when choosing an effective study method is supported in the literature stemming from a focus on external rewards (grades) or a lack of knowledge on what method works best (Weissgerber, 2019). When students` feedback

is solely based on short-term test results, blocked practice, characterized by single study sessions, may seem effective. However, the optimal learning strategy depends on the desired outcome. To pass a test, blocked practice might suffice. In contrast, for goals such as L2 communication, diverse strategies are necessary. This section explores five specific DDs in the context of SLA, providing insights for tailoring effective learning methods to specific tasks and learners.

2.1.1 Spacing

Spacing refers to the reduced drop-off in recall observed when the content being learned is reviewed regularly, as opposed to content being learned in a blocked fashion, where learning is focused at a one-time point or a single learning session. Based on initial findings by Ebbinghaus (1885), as described in Didau (2015), evidence supports the use of spacing in a range of learning contexts. More recent studies have reproduced Ebbinghaus` findings (e.g., Murre & Dros, 2015) and have shown a benefit in different learning contexts (e.g., Cepeda et al., 2008; Rohner & Taylor, 2007). Of relevance here, including L2 vocabulary learning, spacing has been shown to improve vocabulary retention for Japanese learners of English at the university level (Nakata, 2015), and younger (5th grade) Farsi learners (Lotfolahi & Salehi, 2017).

According to Brown, Roediger, and McDaniel (2014), spaced practice is more effective than blocked practice for promoting long-term learning. By increasing the number of intervals and time between review sessions, learning has been shown to be durable. In contrast, blocked practice lacks the necessary time for long-term learning to occur. Spacing can initially lead to forgetting and difficulty in retrieving studied content. However, this struggle ultimately

strengthens memory connections by creating multiple access routes to learned material (Bjork, 1994; Bjork & Bjork, 2011; Persellin & Daniels, 2018). This difficulty in retrieval is considered a desirable difficulty because it enhances recall ability.

2.1.2 Interleaving

Interleaving, unlike blocked practice that focuses on a single skill, involves the concurrent practice of multiple related skills. This requires the learner to discern the appropriate skills for each task, leading to a slowing down of the learning process (e.g., studying present tense and past tense verbs together). This benefit of interleaving stems from increased cognitive processing of the implementation of the correct skill (Bjork & Kroll, 2015). Research shows that if the concepts, methods, or skills interleaved are too dissimilar, the increased difficulty from discriminating would be less likely to be present; therefore, the desired improvement may be less likely to occur (Agarwal & Bain, 2019; Carvalho & Goldstone, 2015; Hausman & Kornell, 2014).

In an authentic situation (e.g., communicating in an L2), multiple skills that are similar are likely to be used. One reason for this is that the learner may have to completely switch from one method, such as using verbs in an L2, to a disconnected task, such as understanding a mathematical concept. This could lead to an attention carryover effect known as attention residue (Leroy, 2009). Attention residue can cause a severe delay when moving between tasks because the residue from the previous task may place too much burden on the learner`s working memory. This effect on working memory is discussed in detail in Sections 2.1.5.1 and 2.5.1.

The risk of interleaving becoming an undesirable difficulty when implemented incorrectly must be considered. Therefore, understanding the correct balance of differentiation between skills is needed for effective attention to task while maintaining enough difference to require the learner to differentiate between skills. In addition to an undesirable difficulty from attention residue, interleaving skills and content that is excessively dissimilar may result in a lessening or a loss of the discrimination between content and skills in the study session (Agarwal & Bain, 2019).

Research on interleaving in mathematics reveals significant learning improvements. Rohrer, Dedrick, and Stershic (2015) found a 25% improvement in one-day post-tests and three-fold (76%) improvement in one-month post-tests among seventh-graders. A larger study by Dedrick and Stershic (2015) confirmed these findings, showing nearly double the improvement with interleaving compared with blocked practice in a real-world classroom setting.

Studies on interleaving in the L2 language learning classroom include grammar exercises and vocabulary learning. One study on ELLs, L1 Japanese, compared English grammar studying of blocked with interleaving practice. Nakata and Suzuki (2019) examined 115 Japanese ELLs at a university in Japan studying five grammar structures. The study found that interleaving led to more errors during study sessions but resulted in greater improvement in one-week post-test results, particularly among learners with weaker grammar skills. The researchers concluded that interleaving could enhance grammar structure practice in L2, although prior grammar knowledge and test-taking skills might influence its efficacy. They suggested that some students might initially require blocked practice before transitioning to interleaving. Schneider, Healy, and Bourne (2002) studied L2 vocabulary learning through

interleaving and blocked practice among L1 English to L2 French learners. While immediate post-tests showed benefits of blocked practice, one-week post-tests revealed little difference between the two methods. Further studies (Nakata & Suzuki, 2019; Pan et al., 2019) found interleaving more effective than blocked practice in delayed post-tests for L2 grammar skills and verb conjugation.

Interleaving in L2 contexts may not be universally effective. Carpenter and Mueller (2013) found that, unlike grammar and vocabulary, L2 pronunciation benefited more from blocked studying than interleaved practice. This difference is attributed to the processing requirements of the activity and the focus on similarity rather than contrasting skills. However, a meta-review by Dunlosky et al. (2013) found interleaving to be particularly effective for problem-solving learning, as it requires skill-switching at each step, potentially providing more effective practice than blocked sessions.

2.1.3 Retrieval

Retrieval, which is also referred to as the testing effect, utilizes active recall of the material studied. This can include content such as facts, definitions, or general comprehension. This DD has been shown to improve long-term memory and recall ability by strengthening neural pathways through the process of recalling specific content (Bjork, 1994; Roediger & Butler, 2011). Retrieval has been well-researched and has demonstrated improvement over simply rereading text or study notes (Kirschner & Hendrick, 2020; Roediger & Karpicke, 2006). Multiple benefits are claimed from using retrieval in learning, including improved long-term retention, knowledge transfer, and the ability to identify gaps in knowledge (i.e., avoiding false fluencies) (Agarwal & Bain, 2019).

Studies have shown the importance of retrieval practice in learning and memory. One area of relevance to learning is that of reading, where retrieval has been found to be 30% more effective than re-reading over different study sessions (Roediger & Karpicke, 2006); nevertheless, students consider re-reading to be a more effective strategy (Agarwal & Bain, 2019; Roediger & Karpicke, 2006). This suggests that students may choose less effective learning strategies because the method (e.g., re-reading) feels more comfortable or less difficult than trying to retrieve content.

McDaniel et al. (2011) looked at retrieval practice learning in the L1 context by comparing studied content that was reviewed three or more times but did not receive retrieval practice with content that was quizzed (retrieval practice) through low-stakes quizzes with feedback. The participants who received regular low-stakes retrieval quizzes had an average grade of an "A-" while those who studied multiple times without retrieval practice quizzes had an average grade of a "C+." Other studies have also found positive results through repeated retrieval (e.g., Carpenter, Pashler, Wixted, & Vul, 2008; Karpicke, 2016; Karpicke & Grimaldi, 2012).

These findings are further supported in various disciplines by retrieval practice meta-analyses. One meta-analysis by Adesope, Trevisan, and Sundararajan (2017) looked at 118 research studies and over 15,000 total participants. It concluded that retrieval practice was more effective than most other traditional studying methods (e.g., rereading or reviewing notes). Agarwal, Nunes, and Blunt (2021) performed a systematic review of the literature on retrieval practice looking at 37 studies and over 5,000 total participants. This meta-analysis looked specifically at the classroom settings from elementary school through medical school and

included multiple disciplines. The researchers found retrieval to be broadly effective across disciplines and age levels.

Studies in L2 learning, specifically vocabulary learning, have repeatedly shown retrieval practice to be effective. One study by Li et al. (2022) looking at Chinese (Chinese L1) university students learning French vocabulary through English as the L2, considered by the researchers as French L3, found that retrieval practice enhanced learning over repeated studying of viewing the word pairs (English to French and French to English) in the longer-term recall tests. The researchers also noted that students were more confident that traditional methods (e.g., rote style learning or reading of vocabulary pairs) would be more effective despite the opposite findings in the post-test results, as described in Roediger and Karpicke (2006).

2.1.4 Feedback

Feedback as a DD focuses primarily on correctional and elaborative feedback. Feedback is defined as the evaluation of the efficacy of a specific learning task. It often comes from external sources like instructors or peers, but can also be metacognitive, where learners provide self-feedback (Agarwal & Bain, 2019). The section explores the potential roles of both external and metacognitive feedback as DDs.

Feedback as a DD has two functions: facilitating learning processes by connecting with prior knowledge or schemas and helping the learner with misconceptions (identifying false fluencies). Instructor and peer feedback, as well as metacognitive feedback, are used to address these false fluencies (Agarwal & Bain, 2019; Brown, Roediger, & McDaniel, 2014). Comprehensive and delayed assessments in the form of post-tests and quizzes, have been

shown to be more beneficial in providing precise feedback over immediate assessments, as they allow for the consolidation of studied content into long-term memory. Despite this, some have continued to claim feedback as a DD is effective soon after studying or instruction on the target material (e.g., Evans et al., 2010; Lee, 2013). The rationale for immediate feedback is to help avoid repeating mistakes or fossilizing incorrect information (Weissgerber, 2019). However, support for delayed over immediate feedback through in-class activities such as low-stakes quizzes is strong (Agarwal & Bain, 2019; McDaniel, Wildman, & Anderson, 2012). In this in-class context, delayed feedback (e.g., low-stakes quizzes delayed by a day or two following the learning task) may be preferred to immediate feedback as immediate feedback can play the role of a crutch limiting the inner thought processing of self-feedback; therefore, delayed feedback may lead to better long-term efficacy in a classroom setting (Brown, Roediger, & McDaniel, 2014).

Elaborative feedback utilized as a DD, can be external or internal and is described as a comprehensive approach by Agarwal and Bain (2019). It often forms part of generative learning, which is discussed in detail in Section 2.3.5 and has been shown to enhance transfer learning, motivation, and understanding of complex tasks (Finn, Thomas, & Rawson, 2018). A study by Wang, Gong, Xu, and Hu (2019) on 104 Chinese university students found that elaborative feedback significantly reduced the negative impact of task complexity. However, the effectiveness of elaborative feedback in second language (L2) learning is inconsistent. Bitchener (2012) suggests that while it may not differ from single-item feedback for advanced readers, it could be more beneficial for L2 learners with lower reading abilities due to the opportunity to form new connections with existing knowledge. Brown (2017) supports this in a study on 113 Emarati English Language Learners (ELLs) at the university level, comparing the

effects of elaborative and corrective feedback on English reading ability. The study found that while elaborative feedback did not significantly improve the performance of intermediate and advanced English readers, it did benefit less proficient readers. Brown concluded that for effective use of elaborative feedback in L2 reading classrooms, it should be targeted to the appropriate learner level with proper scaffolding.

Research on corrective feedback in writing, in both L1 and L2 contexts, has shown the potential to enhance future writing skills (Tsao et al., 2017). A study by Yu, Jiang, and Zhou (2020) on 1,190 Chinese university students learning English revealed that certain feedback types improved motivation. The feedback types included scoring, process-oriented, written corrective, peer and self-feedback, and expressive feedback. The study found scoring, peer and self-feedback, and expressive feedback positively influenced motivation, while written corrective and process-oriented feedback had a negative impact, with process-oriented feedback increasing anxiety and failure avoidance. Feedback as a DD, though not as clearly defined as other DDs, can yield significant effectiveness when employed appropriately (Agarwal & Bain, 2019). Though research has shown that feedback may not always be advantageous, the literature supports elaborative feedback that is delayed as being the most consistent in maximizing effectiveness. Immediate feedback, on the other hand, can lead to dependency for learners, which potentially limits the advantages associated with feedback as a DD (Brown, Roediger, & McDaniel, 2014).

Feedback as a DD is particularly evident in writing, where students must revise and rewrite to improve their skills (Brown, 2012; Hyland & Hyland, 2019). Clear explanations of mistakes from instructors are crucial (Wilson, 2009), and without them, students may focus solely on grades (Kohn, 2011; 2013). Elaborative feedback encourages deeper contemplation

and can be used in future tasks (Nicol & Macfarlane-Dick, 2006; Wilson, 2009). Omitting grades may lead students to consider feedback more thoroughly (Kohn, 1993). Teachers can further prompt deep thinking and self-feedback by asking questions during assessments (Agarwal & Bain, 2019). This approach of self-feedback may be effective for correcting mistakes, but not errors that require instructional guidance (Ellis, 1997). The literature supports feedback as a DD being effective when appropriately applied, with delayed elaborative feedback being most beneficial (Agarwal & Bain, 2019; Brown, Roediger, & McDaniel, 2014).

2.1.5 Generation

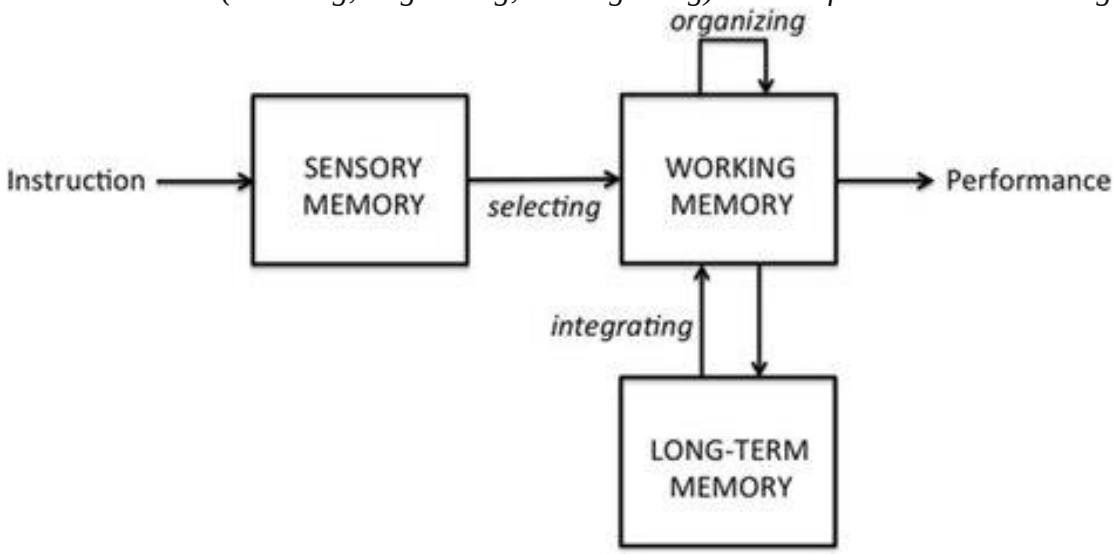
Generation, also known as the generation effect or elaboration, is a pedagogical approach that emphasizes the importance of deep engagement with the material being studied. Fiorella and Mayer (2015) explain that generation falls under the constructivist learning theory, requiring students to construct knowledge during the learning process using their schema to build (construct) upon. Generation learning originated with Piaget (1926) and Bartlett (1932) (as cited in Fiorella and Mayer, 2015), when it was considered a construction connecting new experiences with prior knowledge instead of a more rote-based learning method around memorization. In the modern era, generation is built upon Wittrock`s work connecting generative learning to more modern theories of learning and practice methods in how students learn and perform during complex cognitive tasks (Welle-Donker et al., 2022). Wittrock argued that the brain builds models when learning and does not learn through passive practice (Wittrock, 1992). The core of Wittrock`s model is making connections with prior learning (transfer) (Fiorella and Mayer, 2015). While much of Wittrock`s generative learning research

and his models were concentrated on reading comprehension in the L1 context, it was also expanded to other areas of learning, such as mathematics and science (Wittrock, 1974a).

The generation approach requires learners to actively generate meaning, examples, or content related to the material, which may lead to a more thorough understanding of the subject matter. Since it is primarily concerned with learning transfer, it is used for deeper understanding and long-term learning through making connections with new content, leading to new learning (Enser and Enser, 2020; Fiorella and Mayer, 2015). Furthermore, it enables the learner to personalize the material being studied by creating (generating) novel output that directly relates to the learner`s experiences. While this process may slow down the rate of studying, it is argued to result in more effective learning outcomes (Fiorella and Mayer, 2015; Wittrock, 1992). According to Enser and Enser (2020) and Fiorella and Mayer (2015), generation can be understood as a process by which learners make sense of new information by relating or making connections to their existing schemas.

The premise of generative learning is that learning happens through the use of the appropriate cognitive process as new information is received (Fiorella & Mayer, 2015). A model used to explain the stages of generative learning is the SOI Model of Generative Learning, where S = selecting (or concentrating) on relevant material, O = organizing the material coherently, and I = integrating connections between cognitive structures material held in long-term memory (e.g., the learner`s schema). Figure 2.1 summarizes the SOI model of generative learning.

Figure 2.1.: The SOI (Selecting, Organizing, & Integrating) Model of Generative Learning



Source: Fiorella and Mayer, 2015

The Generative Learning SOI model represents a framework that describes and improves generative learning processes. It accounts for the cognitive mechanisms that underlie successful learning and provides insights into how learners optimize their cognitive resources during the learning process. This works as metacognitive feedback for the learner as described in Section 2.5.3. The initial step in the model is selecting external material that is received from initial instruction (e.g., from an instructor or textbook) through sensory memory. Once received, the selecting stage takes place. Selection requires the learner to choose and process specific sensory input by holding it in working memory. While in working memory, the learner organizes and generates (constructs) a mental representation of the selected material. If needed, the learner can call upon long-term memory knowledge or experiences (schemas) to integrate with the incoming material held in working memory, which enables the learner to make a mental representation if they have sufficient L2 linguistic ability and schema. This newly

generated knowledge can be used to solve problems or make new long-term memory stores (Fiorella and Mayer, 2015).

In a practical learning context, the cognitive processes can be seen in the following example of a summary writing task, which requires the restating of core ideas from a text, video, or lesson into the learner`s own words. The process is initiated with the instruction material through sensory memory. At this point, the learner selects key information, which is held in working memory, and organizes it. From here, schemas from long-term memory stores (e.g., language skills and content knowledge) are integrated into the organization to help the learner understand and make a mental representation of the content being summarized. The final step is the novel-generated content described here as performance in the S.O.I. model presented in Figure 2.1. It is important to note that performance, as described in the S.O.I. model, represents generated content and is not the same as performance discussed in Section 1.3.1.

Generation differs from other DDs in that it takes a more active role in generating knowledge, while other DDs add difficulty by forcing the learner to work through the same material originally studied. For example, when learning vocabulary items and implementing spacing, the time between study sessions increases, but the actual content is not necessarily altered. Generation would require the learner to produce novel sentences and connect the vocabulary to their existing schema. According to Fiorella and Mayer (2015), spacing and interleaving are concerned with the scheduling of practice rather than the actual practice method, while generation is focused on the method. Generation also differs from retrieval, where retrieval focused on retention and how much is remembered, whereas generation is concerned with transferring knowledge, which illustrates how well one understands.

Active engagement with content through generative learning methods has demonstrated an improvement in recall. Chi et al. (1989) found that students who generate active examples through self-explanation (a form of elaboration) of sample problems in an L1 context (studying physics problems in university) were more likely to move from novice to expert learners of the material. The researchers contend that students who generate their own explanations are more likely to find gaps in their knowledge and make corrections. Moreover, they found that students who generated self-explanations were less likely to reread problem sets and explanations. In contrast, less cognitively active study strategies (e.g., highlighting and rereading) show less consistent benefits. Dunlosky, Rawson, Marsh, Nathan, and Willingham (2013) conducted a study technique analysis of ten methods, looking at their utility in studying. Of the ten techniques, highlighting (underlining) and rereading, while not considered to affect learning negatively, were rated by the researchers as having low utility when compared to other techniques, such as retrieval or interleaving. The researchers contend that the methods lead students not to implement more engaging or productive strategies. Surprisingly, these are considered to be two of the most common techniques in studying, yet some of the least effective. Karpicke, Butler, and Roediger (2009) argue that this is due to their passive nature and the false fluency students get from short-lived familiarity derived from the techniques. The relative lack of consistent efficacy of these passive strategies compared to findings of more active approaches, such as the generation of information, has been supported by studies across disciplines among L1 contexts (Bertsch, Pesta, Wisconsis, & McDaniel, 2007; Richland, Bjork, Finley, & Linn, 2005; Zormpa, Brehm, Hoedemaker, & Meyer, 2019).

In SLA, learners use generation to improve their understanding and consolidation of concepts, vocabulary items, and grammar structures. They do this by creating their own

examples, describing authentic or imaginary situations, and asking original questions after reading. Fiorella and Mayer (2015) identify various activities that promote understanding through generation, such as summarizing, mapping, drawing, self-testing, self-explaining, teaching, and enacting. Generative learning involves constructing mental representations by integrating prior experiences and knowledge with the topic being studied during the generation activity. For example, the idiom "to beat a dead horse" generates a mental picture or story about something that the learner can make unique associations with. While not specific or limited to idioms, connection to prior knowledge helps with faster and more durable learning (Clarke, Ayres, & Sweller, 2005; Dong, Jong, & King, 2020; Hattie, 2012; Webb & Chang, 2015). In L2 classrooms, generating dialogues using target grammar or vocabulary is a common task. Self-regulated learners take on the role of both learner and designer of their own learning process, monitoring their cognition, motivation, and progress (Kornell & Bjork, 2007). The use of generation as a DD involves creativity and self-awareness, connecting new learning with previously acquired knowledge to prevent false fluency. The use of generation is highly flexible in L2 classrooms, and the next section will discuss specific applications of this DD in L2 vocabulary acquisition.

2.1.5.1 Detriments of Generation in L2 Learning

Much research has supported generation in learning in various disciplines in an L1 context. However, generative learning in the L2 context is not as well supported. Moreover, there is research that has found generation use to be detrimental. One such area is in L2 vocabulary learning. Barcroft (2009) conducted experiments looking at L2 vocabulary retention among adult English language learning university students. Using expository texts at two

levels, one with easier vocabulary and one with more difficult vocabulary, Barcroft had participants generate novel meaning through synonyms of vocabulary items in the texts. This led to lower scores from generation use with synonyms on L2 vocabulary recall. In a follow-up study, Barcroft (2015) found that a generation task may enhance vocabulary learning if a clear definition is initially provided and there is an opportunity to use retrieval tasks of the vocabulary in context during reading activities. In this second study, the level of generation required was reduced. Instead of novel answers, students in the generation group used retrieval tasks (e.g., retrieval through fill-in blanks for meaning) for target words that were introduced prior to the reading. The goal was to increase vocabulary learning during the reading process and to increase the focus on reading for meaning. The previous (Barcroft, 2009) study did not use the retrieval aspect, just generation. The implementation of retrieval with generation led to a 51% improvement over the non-retrieval group. Barcroft concluded that the use of retrieval was responsible for the improvement since it reduced the reliance on generation.

The efficacy of generation in L2 contexts is explained by Richland et al. (2005) as showing some positive effects in controlled laboratory settings, but when implemented in a classroom setting, the data is unclear. Richland et al. (2005) and Dunlosky (2013) contend that the lack of clear benefits combined with the amount of effort required when using generation calls into question whether its use is optimal. Generation encompasses a wide range of factors and is interconnected with other DDs, thereby complicating the assessment of its effectiveness. The use in conjunction with retrieval may further confound the efficacy of generation. Other variables may also play a role and influence the efficacy of generation, such as the amount of working memory resources the learner requires.

2.2 Counterargument to Desirable Difficulties

2.2.1 Introduction to Cognitive Load Theory

As discussed in this thesis, DDs involve introducing more challenging material with the aim of promoting deeper understanding and long-term retention. On the other hand, Cognitive Load Theory (CLT) argues that overloading working memory during the learning process can cause a negative effect leading to a loss of information held in working memory stores (Bates, 2019). CLT looks to avoid or limit unnecessary complexity or difficulty during the learning process, which can lead to cognitive overload. Cognitive load refers to the amount of information a learner can hold in their working memory, which is limited (Sweller, 1988). When the attention demands exceed the limits of working memory, it can lead to cognitive overload, which occurs when the processing capacity is pushed beyond its limits, making new learning ineffective.

Working memory differs from short-term memory, as defined in this thesis, in that it is considered responsible for temporarily holding and manipulating information during mental tasks. It plays a crucial role in tasks that require immediate processing, such as language comprehension. Working memory allows the learner to hold information in their mind while they use it to perform various mental activities. Conversely, short-term memory is used in the short-term storage of content but is not involved in the manipulation of information. Therefore, working memory is vital in L2 learning, enabling the temporary storage and processing of L2 input. Unlike long-term memory, it is limited in capacity, which poses challenges; however, educators can alleviate some of these challenges by breaking down complexity and offering

scaffolded practice (Sweller, 2011). CLT further suggests that reducing the load on a person's working memory can free up more cognitive resources, improving their ability to transfer new information to their long-term memory, also known as encoding new information.

CLT is based on the concept of Human Cognitive Architecture (HCA) (Geary, 2008). Geary's research has identified two types of knowledge: primary and secondary. Primary knowledge refers to knowledge or skills people have evolved to do, such as walking or eating. This thesis will focus on secondary knowledge, which refers to intentionally learned content, such as L2 vocabulary exercises. The processing, storage, and use of this content is a key concern of CLT (Kirschner et al., 2011). CLT posits two types of activities: those with higher and lower levels of element interactivity (Lovell, 2020), which determine the level of interactivity and, thus, the level of cognitive load. Higher element activities are more complex and require more cognitive resources. Complexity, rather than difficulty, determines whether an activity is a higher or lower element. While an L2 vocabulary exercise with 25 new terms requiring L2 to L1 translation may be challenging, it is still a lower-element activity since it is not complex. Additionally, an L2 learning task that involves pronouncing vocabulary items, conjugating verbs, and reading comprehension would be at a higher level of complexity if the learner is not yet proficient. However, the same task without speaking or conjugating tasks would be considered a lower element. CLT argues that the more complex the activity, the more likely that working memory will be overwhelmed (Bates, 2019).

CLT suggests that complexity requires the learner to hold multiple competing learning points in their working memory at once. Therefore, if the learner already has a schema for some

of the content or aspects of the content, the load required to hold in working memory would likely be reduced, leading to a lower level of complexity. For example, if an ELL has not yet learned the alphabet, learning English (L2) vocabulary by writing words would likely be too complex. In this case, learning the English alphabet first would be the schema that would lessen the complexity of writing L2 vocabulary. Garnett (2020) explains that prior knowledge (i.e., schemas) and understanding are directly related because knowledge itself is the collection and connections between learned content (e.g., concepts, facts, ideas). This can be seen in chunking: the grouping and/or organizing of information into individual units (Lovell, 2020; Paas & van Merriënboer, 2020). By combining multiple pieces of information into single units, less complexity may occur, which CLT argues reduces the demand on working memory during the encoding process (Garnett, 2020; Lovell, 2020).

2.2.2 Cognitive Load Theory and Desirable Difficulties

Some proponents of CLT, such as Cadaret and Yates (2021), and Chen et al. (2015, 2016, 2018) have investigated the relationship between DD and CLT theories with respect to complexity and found when complexity is high, the addition of DDs can lead to cognitive overload during studying. However, the counterargument to DDs primarily applies to higher-level activities, as the use of DDs in lower-element activities often shows a benefit, which is accepted by CLT proponents as well (Bjork, Dunlosky & Kornell, 2013; Bjork & Kroll, 2015; Chen et al, 2016; 2018).

There appears to be a contradiction between the theory behind DDs and that of CLT when it comes to learning in some contexts (Chen et al., 2018; Nihalani & Robinson, 2022).

One reason for this could be the emphasis on the long-term benefits of DD over short-term gains in performance (Soderstrom & Bjork, 2015). DDs seem to work best for learning that does not require many elements, while more complex activities like expository tasks show less benefit. It is important to note that there is a lack of research on DD use in expository texts to make confident conclusions (Brunmair & Richter, 2019). Despite the findings of CLT proponents that DDs may lead to cognitive overload in more complex learning, others have found that the use of DDs (e.g., in the case of interleaving), is more effective for complex high-element content learning due to similarities between categories (Carvalho & Goldstone, 2017; Kornell & Bjork, 2008).

2.2.3 Generation and Cognitive Load

Generation naturally incorporates multiple elements, which encourages learners to process and produce at a deeper level than the presented content alone (Weissgerber, 2019). Therefore, it is likely that the use of generation leads to an increase in complexity, which may have a negative effect of overloading cognitive load. Consequently, the question arises as to whether learners benefit from using generation and, if so, to what extent. One perspective is that learners with lower ability or limited foundational knowledge in a particular task or content of the discipline (e.g., target L2) may not show significant efficacy compared to those with higher foundational knowledge (Sweller, 2011). Weissgerber (2019), however, suggests that it is not the generation effect that affects learning efficacy but rather the process stimulated by the generative task. In simple terms, learners lacking foundational knowledge may not benefit from generating activities that require deeper processing. This aligns with the CLT proponents`s arguments who advocate for limiting complexity in learning. Therefore, according to some

CLT researchers, learners must have a competent level of foundational knowledge to fully take advantage of generation (Chen, Kalyuga & Sweller, 2015; 2016). While generation may cause an increase in complexity among learners in different contexts, learners in an L2 context may have a further level of inherent complexity, increasing the number of elements (Harrington & Sawyer, 1992; Vandergrift, 2007).

This thesis investigates the role of generation as a DD in multiple L2 contexts. The argument that generation tasks may incorporate more complexity from the multiple elements involved during the learning process than other DDs (e.g., spacing and retrieval) (Weissgerber, 2019), makes it necessary to include CLT as a possible counterpoint to generation. Learners inherently have more to process when processing takes place in L2; therefore, this may have a negative effect on learning from generation tasks, which may differ from other findings in the L1 context (e.g., DeWinstanley & Bjork, 2004; McDaniel, Waddill, & Einstein, 1988; Wittrock, 1974b; 1989). This thesis aims to investigate and help bridge the gaps in what the research shows regarding the utilization of generation as a DD in L1 contexts to the L2 contexts by looking at vocabulary learning, reading comprehension, and academic writing in the L2 classroom.

2.3 Second Language Vocabulary Learning

Vocabulary is a key aspect for second language learners to improve their skills in that they must learn a wide range of words to be able to express themselves clearly, both orally and in writing, and is considered to be an important focal point in acquiring L2 (Nation, 2001). Hamada and Koda (2008) express the importance even more strongly by claiming that

acquiring vocabulary is the most crucial element of learning a second language. Vocabulary acquisition is essential from the start of learning a new language, and it often begins with basic greetings in the classroom. While early exposure to vocabulary can help with basic understanding, a deeper level of understanding beyond basic meaning, such as the ability to use the target word in different contexts and forms. These forms can include meaning, grammatical functions, parts of speech, register, word associations, and collocations (Nation, 2020), which are required to create content and share knowledge, which is necessary for progress in the target language. L2 vocabulary learning is influenced by both intralexical and interlexical factors (Schmitt & Schmitt, 2020). Interlexical factors are related to the impact of the learner's L1 on the acquisition of the target L2 vocabulary. This includes the congruency between L1 and L2 words or phrases, and the common grouping or pairing of certain words in use together (e.g., collocations) (Peters, 2016). Conversely, intralexical factors are inherent to the words themselves. These include the word length and the type of collocation, which refers to the relationship between the collocate and the node. Both these factors can influence the ease or difficulty with which a learner acquires the target L2 vocabulary (Schmitt & Schmitt, 2020; Webb, 2020).

Word knowledge encompasses not only understanding the word`s form but also the likelihood of encountering it in speech or writing, contextual constraints on its usage, syntactic patterns associated with it, and awareness of ongoing vocabulary growth (Wen & Naim, 2023). Knowledge of L2 vocabulary is multidimensional in that it consists of both breadth and depth. At the foundational level, breadth refers to the linear or single-factor aspects of word knowledge.

The vocabulary breadth encompasses the number of words known (Webb, 2020), form (e.g., sound, spelling), and meaning (e.g., word associations) (Nation, 2001). The necessary breadth of vocabulary may vary depending on the learner's needs. For example, those aiming to understand written texts may require a larger vocabulary of around 8,000-9,000 word families in English, while a verbal comprehension level of 98% can be achieved with 6,000-7,000 word families (Nation, 2006). In contrast, some experts argue that an adequate (basic general understanding) listening comprehension level can be achieved with as few as 2,000-3,000 word families (Van Zeeland & Schmitt, 2012). This difference in required vocabulary illustrates a divide between written and verbal communication comprehension in L2.

L2 vocabulary depth refers to the level at which the vocabulary is known. Depth knowledge encompasses word meanings, collocations, semantic relationships, and syntactic patterning (Cobb, 1999; Webb, 2020). Nation & Webb (2011) and Schmitt and Schmitt (2020) emphasize the importance of both breadth and depth in vocabulary knowledge for effective communication in the target L2 as breadth assists the learner to understand in more contexts, while depth enables the learner to understand the nuances or word usage.

2.3.1 Memory and Vocabulary

Memory is a core component in the acquisition of L2 vocabulary. Although the study of memory is extensive, this section will concentrate on second language vocabulary acquisition among classroom learners, with a particular emphasis on instructional methods. The aim of this section is to lay a foundation for the processes by which information is transferred from initial exposure (input) to a state where it can be recalled and utilized weeks later. This is referred to as the encoding process. Three types of memory are discussed in this thesis: sensory, short-term

(working), and long-term memory, with the primary emphasis being on short-term and long-term memory, derived from Baddeley, Eysenck, and Anderson (2014). Sensory memory is the initial store of information that comes from the senses (e.g., sight, touch, etc.) and lasts up to a few seconds. Short-term memory has multiple definitions, but in this thesis, short-term memory is defined as the retention of information (e.g., a vocabulary item) for a period ranging from several minutes to a few days. In many contexts, working memory is considered interchangeably with short-term memory. However, for the purposes of this thesis, working memory is considered different from short-term memory in that it refers to the amount of memory immediately available for new learning during the encoding process as described in Section 2.2.1. Within the context of this thesis, long-term learning is attained when stored content can be recollected three weeks or more subsequent to the last exposure.

2.3.2 Overview of Explicit & Implicit Vocabulary Learning

When learning L2 vocabulary, students and instructors implement strategies that fall into one of two general techniques: explicit and implicit. Explicit or intentional vocabulary learning refers to the conscious and planned acquisition of targeted L2 vocabulary. Conversely, implicit or incidental vocabulary learning is unconscious and unplanned vocabulary acquisition. The use of DDs in L2 vocabulary learning inherently incorporates explicit over implicit learning techniques as explained in the following sections. In this thesis, explicit and intentional vocabulary learning are used interchangeably, as are implicit and incidental vocabulary learning unless otherwise noted.

2.3.3 Explicit and Intentional Vocabulary Learning

Explicit learning of L2 vocabulary is often associated with studying terms out of context through word lists; however, this is not always the case, as the act of looking words up in a dictionary or taking notes on vocabulary words during reading exercises are intentional learning techniques in context (Nezhad, Moghali, & Soori, 2015). Explicit learning can aid learners in gaining a comprehensive understanding of word ranges and uses and applying them in different contexts. Strategic approaches such as contextual learning, dictionary use, rote rehearsal, and vocabulary in use involve an intentional and iterative process that may be optimal for effectiveness and efficiency (Gu, 2003). Aspects of whether a specific learning task requires breadth or depth of vocabulary knowledge, in addition to what background the learner has on the topic or related vocabulary, play a role in the efficacy of L2 vocabulary acquisition (Gu, 2003). This is illustrated by intentional learning`s flexibility through scaffolding of targeted L2 vocabulary in a given task. The flexibility provided by intentional methods, such as word lists and flashcards, may be particularly significant in diversifying study techniques used to internalize new vocabulary (Hill & Laufer, 2003; Schmitt & Schmitt, 2020).

Supporters of intentional learning highlight its potential to achieve better comprehension and retention of vocabulary compared to incidental learning, as explicit vocabulary learning has been found to improve L2 proficiency overall, as illustrated by Yousefi and Biria (2018). Yousefi and Biria conducted a meta-analysis of 16 studies on L2 vocabulary teaching and found that L2 vocabulary instruction is an effective and indispensable part of the L2 curriculum for efficient learning. This is further supported as explicit instruction in vocabulary has been shown to speed up the rate of acquisition and enhance retention (Schmitt

& Schmitt, 2020). Yousefi and Biria (2018) suggest that intentional learning of L2 vocabulary results in a deeper grasp of the words, improving their retention.

2.3.4 Intentional Vocabulary Learning Strategies with DDs

In intentional learning of L2 vocabulary, the use of DDs commonly plays an important role in the strategies used by learners. Various studies have revealed that intentional learning through flashcards, wordlists, and vocabulary note-taking can have positive results (Elgort & Nation, 2010). Furthermore, flashcards may be the most common L2 vocabulary learning strategy used by students both in the classroom and among self-directed learners. Wissman, Rawson, and Pyc (2012) found that nearly 67% of university students use a flashcard strategy, and Schmitt (1997) found that more than 50% of Japanese junior high school students studying EFL use flashcard strategies. The use of flashcards is particularly relevant to this thesis as their use commonly implements the desirable difficulties of spacing, retrieval, and interleaving.

A study conducted by Hung (2015) investigated paper-based flashcards used in the university EFL classroom with peer practice and digital flashcards with self-study practice. Word knowledge, form, and part of speech were assessed through L1 to L2 translation recall through two-week post-tests, which found that digital flashcards through a self-study condition outperformed peer and group work where communication was involved. Additionally, a meta-analysis of 22 studies on intentional vocabulary learning through flashcards, wordlists, vocabulary note-taking, and fill-in answers conducted by Webb, Yanagisawa, and Uchihara (2020) found that immediate assessment showed a relatively large improvement, yet the gains were fleeting as delayed assessments were found to be more modest. Nevertheless, flashcards and wordlists were found in their analysis to be particularly effective in learning the form-

meaning of words. Other studies on L2 vocabulary learning support the findings discussed in this section on intentional learning using spacing and retrieval (e.g., Barcroft, 2007; Karpicke & Roediger, 2008; Nakata, 2011; 2017).

The availability of multiple study methods (e.g., use of wordlists and flashcards) may be particularly significant in intentional learning (Hill & Laufer, 2003; Schmitt & Schmitt, 2020). By diversifying the techniques used, learners are presented with a range of approaches to internalizing new vocabulary. This diversification of methods, in addition to encountering items in different ways and contexts, ultimately may produce a more robust vocabulary repertoire, as different learners may respond better to specific methods (Schmitt & Schmitt, 2020), leading to improved retention of the target vocabulary (Elgort, 2011). Furthermore, the concentration on specific vocabulary items during intentional learning exposes learners to repeated instances or initial uses of the target words. This exposure may enable learners to recognize these words when encountered incidentally at a later time. Consequently, deliberate attention to particular vocabulary items has been found to both improve initial encoding and contribute to reinforcing vocabulary knowledge during subsequent encounters in a more natural context (Elgort, 2011; Ellis, 1994; Hulstitijn & Laufer, 2001).

2.3.5 Implicit and Incidental Vocabulary Learning

While implicit learning of L1 vocabulary accounts for most of L1 vocabulary acquisition (Webb & Nation, 2017), its role in the acquisition of L2 vocabulary is less clear. As discussed, explicit learning plays a strong role in L2 vocabulary acquisition. Nevertheless, Krashen (1989) argues that adequate exposure to language through listening and reading can lead to sufficient vocabulary acquisition without the need for extrinsic study methods like

flashcards or rote memorization. Frequently used words in the target language may provide enough exposure and variety of uses for learners to develop a strong understanding (Krashen, 1993; 1999).

Nowbakht and Shahnazari (2015) compared incidental L2 vocabulary learning through comprehensible input (see Section 2.2.1) with more active uses of L2 vocabulary along with feedback. A one-week post-vocabulary treatment test showed no significant difference between incidental and intentional treatments regarding output, illustrating that comprehensible input alone may have been the driver of vocabulary acquisition; however, the results did show a benefit of feedback on mistaken vocabulary. This study did not show an improvement in intentional over incidental learning without incorporating teacher feedback. Incidental vocabulary learning is generally thought to play an important role in L2 vocabulary acquisition, but its use as the sole method of learning L2 vocabulary is currently not well supported as the purpose for learning specific vocabulary can differ (Hill & Laufer, 2003; Schmitt & Schmitt, 2020; Webb, 2020). While both implicit and explicit methods have shown efficacy as described, their role as complementary approaches has the most support of efficacy (Schmitt & Schmitt, 2020).

2.4 Second Language Reading Comprehension

2.4.1 Background Introduction

Reading comprehension is a complex process that requires multiple skills that go beyond the scope of this thesis. However, two broad-based skills that encompass the key aspects of reading are language comprehension and vocabulary reading (Oakhill, Cain, & Elbro, 2019). In the context of Japan, English reading has become fundamental in both work

and personal life. In society, the advent of the internet, email, and social media has only increased the need for reading skills. As the world has globalized, the need for learners to develop reading at the professional level in English has increased. DDs may be a tool students can utilize in learning to read in the target L2. The following provides background information on DDs used in reading comprehension tasks. The literature on DDs used in reading comprehension is primarily focused on L1 contexts, and findings in the L2 contexts show mixed results, clearly illustrating the need for more research in L2 contexts. This section covers the following DDs or tasks acting as DDs in reading contexts: disfluency through font manipulation, the use of retrieval, notetaking, and interspersing.

2.4.2. Disfluency as a Desirable Difficulty in Reading Comprehension

While there is less research on the effect of DDs on reading comprehension in the L2 context, one area of research that has been researched is disfluency. Disfluency in reading makes reading the actual text more challenging. It is considered a DD in some situations, while in others, it has been shown to have mixed results in L1 contexts. For instance, Diemand-Yauman, Oppenheimer, and Vaughan (2010) found positive effects in two experiments in an L1 reading context. They illustrated that adding difficulty through the use of a slightly harder to read font increased memory of the reading text in a laboratory setting among university and high school students. The disfluent material presented used 12-point font and 60% grayscale in either Comic Sans MS or *Bodoni MT*. The fluent condition used a larger 16-point of Arial and pure black font in place of the grayscale.

Alternatively, Katzir, Hershko, and Halamish (2013) found that decreased font size and an increase in line length had a negative effect on comprehension amongst second-grade

readers, but fifth-grade readers had improved comprehension for smaller text. The baseline used a 20-point font size with a line length of 4.2 inches and spacing of 2.0, which was the U.S. national standard for second-grade level assessments and textbooks. The disfluency condition was implemented through three different manipulations of the text presented separately. They included a reduction of 20% in font size; an increase in line length by 20%; and a decrease in line spacing by 20%. For the second-grade students, the reduction in font size and the increase in line length showed a significant negative effect on reading comprehension, but the decrease in line spacing did not show an effect. Conversely, the fifth-grade students had a significant positive effect from the decrease in font size, while the other two conditions did not show an effect. The authors contend that children at different stages of development respond differently to the use of disfluency: the more developed the children`s reading stage, the more likely they are to benefit from disfluency. The researchers contend that older readers had already achieved reading efficiency and, therefore, could engage with the text better and improve their comprehension. Conversely, the younger children had not yet reached the stage of reading efficiently and therefore did not experience the same benefits for deeper engagement with the reading. This led to a decrease in reading speed and retention. The younger readers did not have the reading skills to turn the difficulty into a desirable difficulty. This study looked at the L1 context among children, but it may reflect the L2 context where the L2 reader`s reading level, rather than their age, may have a similar effect.

While effects from L1 studies are less clear, findings in the L2 context support a negative effect for disfluency in reading. Dykes and Hauca (2019) conducted a reading comprehension study on Japanese university students to investigate the effect of disfluency through font style. Sans Forgetica (Sans Forgetica) was used as the disfluency font style. It was

developed using the principles of desirable difficulties to enhance recall by adding just enough effort to evoke deeper cognitive processing (Harris, 2018). "Century Schoolbook" font style was the baseline condition. They found that disfluency slowed reading down and led to a negative effect on comprehension. As the treatment disfluency slowed the reading down, but it did not lead to an increase in comprehension, the authors concluded that disfluency through font type among ELLs, or the specific font type of Sans Forgetica, is an undesirable difficulty. They concluded that the impact on reading speed and comprehension is much stronger for L2 readers, which could have resulted in negative results. Although not specific to L2 learners, it is important to note that recent research has questioned using Sans Forgetica as a DD in memory recall. Taylor et al. (2020) conducted four experiments on word pair and reading recall among 882 participants in an L1 context, comparing Sans Forgetica to Arial font style, which is considered a standard font. They found either no effect or a negative effect in each experiment on the use of Sans Forgetica as a DD. Taylor et al. (2020) question the use of Sans Forgetica as a desirable difficulty. They remark that Sans Forgetica is harder to read than standard fonts like Arial, but the difficulty is just that: difficult, but not desirable as Sans Forgetica font showed either no improvement or caused negative effects. They conclude that disfluency does not encourage deeper processing of reading.

Romney (2019) found negative results among Japanese university ELLs, similar to Dykes and Hauca (2019). Romney showed that a disfluent typeface from the class textbook with exploratory readings caused slower reading and lower comprehension scores than a standard typeface for the textbook or a common typeface that is not considered disfluent. The disfluent typeface condition used was Lucida Blackletter, and the two common typefaces were Helvetica and Times New Roman. Helvetica was the baseline condition because it was a

familiar font from the participants` textbook. Times New Roman was used as a condition to compare if any effects come simply from less familiar, but not disfluent, font types. Romney compared the three conditions and found that the most familiar font, Helvetica, produced the fastest reading times and highest mean comprehension scores. The disfluency typeface condition produced the slowest reading times and lowest comprehension scores, leading the author to conclude that disfluency may negatively impact L2 learners' reading. Reading slowly in itself does not indicate a negative impact, but the fact that the learners read more slowly than the control group and still had lower scores indicates a possible negative impact from disfluency in reading speed and comprehension.

While disfluency from font difficulty shows mixed results in an L1 context, the literature clearly shows that its use in the L2 context generally leads to worse outcomes in terms of comprehension accuracy. This may be attributed to the extraneous cognitive load that this difficulty causes, leading it to become an undesirable difficulty for L2 readers. Although the extra difficulty added from these disfluent texts may be a DD in some L1 contexts, the context of reading in L2 itself is an extra difficulty without the desirable effect. Taking these findings into account, finding a way to reduce cognitive load for L2 readers while promoting the addition of generation and deeper thinking could be an effective method of not over-taxing the L2 learner`s working memory while still gaining the benefits of deeper learning explained in the generative process of comprehension model (Wittrock, 1989).

The literature on generation and reading comprehension is largely built on Wittrock`s generative process of comprehension model, which explains how readers actively construct meaning from text by using prior knowledge and making (generating) connections between ideas (Doctorow, Wittrock, & Marks, 1978). According to Wittrock, comprehension is not just

a passive reception of information, but a creative and productive process involving motivation, attention, and memory (1989). The use of active studying techniques, such as generating reading summaries by readers in L1 contexts, is argued by Wittrock (1989; 1992) to lead to improved reading comprehension.

Conversely, Dunlosky (2013) asserts that generative summarization, while effective, requires much effort and time that may be better applied to other strategies depending on the background knowledge that the learner has. He considers generation in reading comprehension to be a promising method that may lead to improved learning outcomes in various areas, including reading comprehension, but thinks that more research is needed before teachers or learners should use it as a primary study method for general use. Abel and Hänze (2019) conducted a study on implementing generation tasks in technical reading to test these approaches. They looked at both comprehension and cognitive load among 199 high school students in an L1 context (German) and found no improvement in immediate performance from the generation treatments. The two treatments were a baseline condition which provided students with the entire text with clauses included, and a generation treatment which gave the same text but left blank spaces for students to choose the correct clause from a choice of four, which was considered a reflection generation task. The researchers did not find an increase in cognitive load assessed from 1) implementing dual tasks (reading the text for comprehension and choosing the correct clause from the choices, with the addition of a randomly appearing elementary math task), 2) recording time on task, and 3) participants self-reporting on a feeling of being overloaded. Nevertheless, they argue that the students were likely overwhelmed by the generation tasks, which may explain the lack of benefit from generation. The researchers assessed cognitive load by comparing conditions where more than one task was required. The

easy math questions that required a quick answer were used to measure cognitive load. If the time on task took longer to answer the math questions, this showed the participants were at a higher cognitive load. The math questions had nothing to do with the Generation task being assessed. Additionally, a post-questionnaire on the effects felt during the task was provided by the participants. Importantly, the conclusions from Wittrock (1989; 1992), Dunlosky (2013), and Abel and Hänze (2019) were all in the L1 context.

In the L2 reading context, there are complex variables that lead to differences in the process from that of L1 contexts (e.g., cultural differences that relate to the schema) (Singhal, 1998). For example, if an L2 learner of Japanese comes from an L1 that lacks background education or knowledge in certain content such as kanji, Chinese written characters, the learner may be at a disadvantage compared to learners from countries that have been exposed to Chinese characters in their L1 (e.g., Taiwan). This would lead the L2 Japanese learner (e.g., a learner from Mexico) to lack the schema (kanji exposure), leading to a need to process content differently, which could put them at a disadvantage. These differences may not be reflected in the literature on generation, as argued in Wittrock`s generative process of comprehension model. L2 readers often encounter difficulties due to their limited vocabulary, grammar, and cultural background knowledge compared to L1 readers (Nation, 2009). The addition of an interaction between L1 and L2 during reading tasks may lead to possible cognitive load issues specific to L2 readers (Harrington & Sawyer, 1992). It is unclear how this presumed increase in pressure on cognitive load may help or hinder the effect of generation in reading tasks.

2.4.3 Retrieval in Reading Comprehension

Studies on retrieval have shown efficacy in L1 contexts in reading and other tasks (e.g., Bjork, 1994; Kirschner & Handrick, 2020; Roediger & Butler, 2011; Roediger & Karpicke, 2006). While generation is not the same as retrieval, it is related and used during the encoding process. Generation requires the learner to take an active role in reading content and using it with their schemas to generate new (novel) information. Persellin and Daniels (2018) explain that this active role through generation can increase retention because generating material requires deeper processing, which helps in encoding. For example, recall is evident through summary generation. When a learner generates a summary of a reading task without the ability to look at the reading text during the summarization, they are forced to retrieve material from memory as part of the process of summarizing content into their own words. This recall is necessary before being able to make connections to their own uniquely generated summary, as without recalling the content of the reading, the learner would not have anything to summarize.

If generation requires retrieval, such as in generating a reading summary, the question of whether or not generation is actually two separate DDs arises. Roelle et al. (2022a) discuss the theoretical connection between retrieval and Generation tasks and how there is a need to clarify the effects of the added difficulty: whether it is generation added to an initial retrieval task, or the opposite of retrieval added to an initial generation task. While inconclusive, there is evidence that generation would be more effective if the retrieval aspect, or some part of it, were deliberately separated from the generating phase. Roelle et al. (2022a) explain that retrieval and generation, while both have efficacy, generally should not be used together. They argue that the use cases of both may not be the same, as retrieval`s function is for encoding content, and generation functions to construct a mental representation following encoding of the content and

for consolidation (Roelle et al., 2022a). They further contend that using two DDs in unison may cause an increase in cognitive load, which was assessed by participants who completed a nine-point rating scale on mental effort and a questionnaire to assess the type of cognitive load (intrinsic versus extrinsic). Intrinsic cognitive load refers to the inherent difficulty required from a specific learning topic, which may result from the learner's existing schema. Extrinsic (also referred to as extraneous) cognitive load addresses outside factors of how the topic is presented or taught. Here, external factors (e.g., instructors or the learning material design) play a controlling factor in how complex or difficult the topic is. They investigated using each of these DDs separately among 158 university students in an L1 context, where one DD would be used for initial encoding of reading an expository text, and the other would be used for consolidation tasks. From comparing both retrieval and generation in both use cases, they showed that retrieval was more effective in encoding and generation was better in consolidation after the initial encoding task took place. Furthermore, when using retrieval before generation, the extraneous cognitive load was reduced compared to using generation for initial encoding before retrieval for consolidation.

Nevertheless, Wittrock and Alesandrini (1990), found that a generation task, which simultaneously incorporated retrieval aspects following the reading of text, was more effective than reading alone. A study of 59 undergraduate university students in an L1 context found that generating summaries after reading a text (generation + retrieval) led to improved imagery of the text and analytic ability in the learning of the text. In contrast, generating analogies (generation only) was effective in analytic ability only, and reading alone without generation showed effectiveness only in text imagery. Moreover, Annis (1985) conducted a study in the L1 context comparing three treatments around reading sections of a university textbook:

sentence summary generation after reading each paragraph, taking notes in the text, and reading only. The students who generated summary sentences scored highest on comprehension, application, and analysis assessment, but those who took notes scored higher on synthesizing and evaluating the reading. She argues that stopping at each paragraph to summarize may have helped in processes such as encoding and making connections to apply the material to real situations, resulting in higher application and analysis scores. As used in these two studies, summary generation of text requires retrieval in tandem with generation. In both Wittrock and Alesandrini (1990) and Annis (1985), generation summaries were found to be more effective for comprehension among L1 readings at the university level, compared to reading alone.

2.4.4 Notetaking in Reading Texts

Taking notes during reading has been shown to be more effective than highlighting or rereading in the L1 context. Willingham (2023) explains that taking notes on generated thoughts from what is being read improves focus on the task. The use of generated notes (created by the reader and not just rewriting the text) has been shown to improve comprehension in L1 reading tasks. Hagen, Braasch, and Braten (2014) investigated the effect of note-taking through argument versus summary writing on comprehension in university undergraduates. They found that simple summarizing (paraphrasing) was less effective on comprehension than notes that were reformulated into more novel-generated argument notes. They also found improvement over simply paraphrasing when generated notes made connections to prior knowledge. It is important to note that summarizing through paraphrasing in note-taking is considered a generative task (Fiorella & Mayer, 2020), but it is less complex than generating an argument. Fiorella and Mayer contend that making connections with prior

knowledge, as is required when generating arguments, provides deeper learning than summarization alone (2020). While generating arguments and making connections from prior learning is argued to be effective in deeper learning in an L1 context, it is unclear if this is the case in an L2 context as much of the note-taking research conducted on L2 learners focuses on lecture-taking notes and not on reading notes for comprehension.

2.4.5 Interspersing Generation in Reading

Interspersing generation tasks within the reading text rather than summarizing the entire text at the end may be effective in comprehension and alleviating extraneous cognitive load. This is through the reduction of the split attention effect of moving between post-reading questions and searching the entire text for where the answers are addressed. This involves placing reflection, summary, or comprehension questions/prompts at certain points in the reading versus at the end of the text. In an L1 context, students can be supported by interspersing summarizing sections over the entire text. Wood (1986) found that university students in an L1 context benefited more from questions placed near the text that the question pertained to compared to questions on pages further away from where the content was located, such as at the end or on separate pages for longer readings. Furthermore, Fiorella and Mayer (2020) propose that this can be used as an alternative method of summarizing to better enable students to locate key information in the text.

In L2 reading, Hung (2009) conducted a study investigating the cognitive load effect of integrating questions into the text versus questions at the end of the text, which Hung considered to be split attention because it required the reader to move back and forth from the questions to the text. Two phases were investigated: a learning phase and a testing phase. The

learning phase required the participants to read the content and answer ten keyword questions within ten minutes (one minute per question). There were two treatments: one treatment embedded keyword questions in the reading at the end of paragraphs where the keyword answers were located, considered to require less extrinsic cognitive load because questions at the end of the reading were believed to cause split attention, as students were required to re-read the text from paragraph to paragraph, looking for specific keyword answers. The second treatment included the same questions, but they were at the end of the entire reading text, considered to cause more extrinsic cognitive load. The split-attention treatment had lower keyword scores than the integrated keyword question treatment. According to Hung, this difference may have been a result of the participants repeatedly jumping from the keyword questions to the text, looking for the answer. This was found to slow down the comprehension process in the testing phase as well. Although the keyword question scores were lower in the split-attention treatment, and the comprehension testing was found to be slower in the testing phase of L2 comprehension questions, which were asked separately from the keyword questions, the comprehension scores themselves did not show a difference from the treatments. The study showed a negative effect of splitting attention (questions coming at the end of the text) on reading comprehension speed, as it took more time and effort for participants to complete reading comprehension questions than for participants in the treatment that integrated the questions into the text. However, the testing phase did not show any difference in answering comprehension questions correctly between treatments.

Al-Shehri and Gitsaki (2010) found that an integrated reading format with comprehension questions inserted directly into the texts facilitated comprehension and reading speed over questions at the end of the reading. This study looked at short online readings

similar in length to those used in the study reported in the rest of this chapter. Participants included 20 young adult ELLs at an intermediate level from more than eight countries studying in Australia. Key variables around interspersing generation in reading may include how often interspersing takes place (e.g., following each paragraph); the depth of generation required; how related to the reading the generation task is; and the types of questions or prompts (e.g., multiple-choice questions or open answer questions).

2.4.6 Summary of Reading and Generation

This section discussed findings on reading and generation in both L1 and L2 contexts. The studies described what has been found in the L1 setting and how it may not be applicable to L2 learners in some contexts, such as disfluency of fonts. However, other areas, such as retrieval, note-taking, and interspersing questions, show possible benefits that need further exploration in the L2 context. Taking these findings into account, finding a way to reduce cognitive load for L2 readers while promoting the addition of generation and deeper thinking may be an effective method of not over-taxing the L2 learners` working memory while still gaining the benefits of deeper learning, as explained in the generative process of comprehension model (Wittrock, 1989).

2.5 Second Language Essay Writing

2.5.1 Background Introduction

Essay writing in L1 contexts is often challenging for university students (Nenotek, Tlonaen, & Manubulu, 2022). In an L2 context, the difficulty increases as learners not only have to consider the content of the writing but must also worry about the mechanics grammar,

punctuation, and proper word choice. To alleviate some of the difficulty, teachers and textbooks used in teaching writing will often provide support to help student writers (Lovell, 2020). This support can affect the demand on cognitive load that student writers must engage with during the writing process. Additionally, instructors should challenge writers and help them become more autonomous in their writing (Hyland, 2019).

In the writing classroom, instructors face the challenging task of ensuring that their students acquire and retain specific skills to a level where they can be applied to new learning (writing). The phenomenon of carryover, or transfer knowledge, plays a crucial role in this regard, as it pertains to prior learning that impacts new learning. Carryover, as defined for the purpose of this thesis, refers to the acquisition of a concept or skill that leads to an enhancement or continuation of learning another future concept, skill, or task. Transfer knowledge is intrinsically linked with the concept of desirable difficulties (DDs) in learning, as defined by Soderstrom and Bjork (2015), which refers to long-term retention that can be applied to novel learning situations. DDs are effective learning strategies that promote the transfer of knowledge to specific skills and critical thinking (Agarwal & Bain, 2019).

Research conducted in L2 contexts on carryover has produced mixed results, with certain studies indicating success (James, 2006) in less cognitively demanding writing tasks and skills (e.g., descriptive writing and grammar), but not in more challenging academic writing tasks (Wubalem, 2021). It is important to note that most carryover studies emphasize on the effect from instructor feedback and the results in writing from one writing class to the next, rather than the transfer from one specific task to another. Furthermore, there is a lack of research on the use of generation and its impact on carryover in the L2 academic writing context, indicating a critical need for further investigations to address the gaps in knowledge.

The following sections describe two foundations in writing education: 1) the cognitive process model of writing, and 2) knowledge-telling and knowledge-transforming models. It then investigates the challenges of generation and the relationship with overtaxing cognitive load in L2 writing. Subsequently, it moves into reflection through generation in L2 academic writing before discussing what the literature offers to alleviate issues of overwhelming students in writing.

2.5.2 Cognitive Process Model of Writing

Flower and Hayes (1981) proposed a cognitive process model of writing that challenges traditional stage models. Their model posits that writing is best understood as a set of distinctive thinking processes that writers orchestrate or organize during the act of composing. These processes include rhetorical (stylistic) decisions, long-term memory, planning, translating, reviewing, and the monitor. In this model, writing is a goal-directed process where writers create a hierarchical network of goals to guide the writing process.

Flower and Hayes' cognitive process model of writing differs from traditional stage models in several ways. Traditional stage models of writing, such as those proposed by Rohman (1965), are linear in the process stages leading to the completion of the writing task. Steps are taken in order, and revision is usually done after completing all steps in the process. Flower and Hayes challenged the linear process models and argued that the cognitive processes in writing have a hierarchical structure and can occur at any time during the composing process. For example, revision is not seen as a separate stage but can occur at any time during the writing process and can lead to new planning or revising of what the writer wants to say. Harmer (2007) explains this through the process wheel, a metaphor of the writing process that

describes how students move through the process non-linearly by focusing on a specific step, viewed as a spoke in the wheel. This process wheel emphasizes that writing is about rewriting through revisiting the writing piece at different stages (spokes) in the wheel in a nonlinear way. The Flower and Hayes` process model is another example considered more exploratory than linear models and accounts for the fact that writing does not always follow a neat step-by-step forward direction and aspects of planning, drafting, revising, and editing interact simultaneously (Hyland, 2019).

The rhetorical problem includes both the rhetorical situation and the audience (e.g., academic) and the writer's own goals in writing. The style or tone of the writing is an example of this rhetorical problem. Long-term memory is used to store knowledge about the writing topic and audience, which also includes knowledge of writing plans and issues around the writing topic. Planning involves the representation of knowledge of the writer and idea generation, while translating involves converting meaning in the writer`s mind to words on the page by making connections with other ideas. It is important to note that translating, as used by Flower and Hayes (1981), does not necessarily refer to translating from L1 to L2, but from taking ideas and turning them into written text. Reviewing is used to unify written text and differs from traditional linear models in that reviewing can take place at any stage in the writing process. The monitor is the writing mechanism that decides what steps should be taken and the stage during which they are implemented. The monitor further serves to guide and choose what thinking process should be used.

2.5.3 Knowledge-telling and Knowledge-transforming Models

One model may not be enough to account for the differences between novice and skilled writers and how they process the complexity of writing (Bereiter and Scardamalia, 1987). Therefore, two separate models, the knowledge-telling and knowledge-transforming models, were developed. Knowledge-telling entails just telling and not changing or creating new knowledge. Novice writers often use this model as they do less planning, have simpler goals, and revise less than skilled writers (Hinkel, 2020). Additionally, the primary concern of novice writers is often with producing content (e.g., primarily the volume of words on a page) (Hyland, 2019). Conversely, skilled writers look to analyze problems, reflect on writing and content, and try to rework ideas, leading to new knowledge creation. This is considered knowledge-transforming. A key difference between these two models is the focus on reflective thought in the knowledge-transforming model for skilled writers, which requires the learner to participate in more cognitively challenging aspects of writing, such as academic writing (Hinkel, 2020). Reflection as part of the writing process can come in the form of feedback (self- and external). This increased challenge may lead to more required resources and, therefore, tax the cognitive load if the writer does not yet have the schema (prior foundational knowledge) to cope with this aspect of writing or be at a disadvantage by having to write in an L2 (Hinkel, 2020). Nevertheless, students need to do more cognitively challenging tasks to develop skills in writing (Hyland, 2019).

While there are two distinct models, writers may not fall distinctly into that of a novice or skilled writer. This is the area where the teacher needs to know how much to challenge the writer to improve their ability while not causing a cognitive overload where learning and writing improvement may not happen. The teacher`s role is that of a facilitator of a writing

process that develops a student`s metacognitive awareness in their writing while moving from that of a novice to a skilled writer (Hyland, 2019). As cognition is a key element of the writing process (Flower & Hayes, 1981), knowing the writer`s thought processes can be key to learning where the student is while moving from novice to skilled writer. Unfortunately, research cannot get into the minds of the writer`s unconscious processing as there are too many variables that affect each task and student (Hyland, 2019).

2.5.4 L1 and L2 Academic Writing

For L2 writers in academic contexts, being able to learn how to write clearly and express what they want to say in the target language can be particularly difficult. Academic writing, such as summary essay writing, can be complex and lead to an overwhelming of cognitive resources for L1 student writers (Kirkland & Saunders, 1991). For L2 writers, the complexity is compounded as they have internal language proficiency constraints (Hinkel, 2020), evident in novice and skilled knowledge models. Students must follow external requirements (e.g., specific style guide formatting), use academic vocabulary that may be unfamiliar (Schmitt et al, 2011), and implement challenging phrasing (e.g., not using first person) in their writing. The compounding of these different aspects alone can overwhelm the cognitive load of the L2 writer (Jagaiah, Howard, & Olinghouse, 2019). In addition, L2 students are often required to reflect on their writing, further leading to possible cognitive overload (McCutchen, 2011) due to metacognitive self-feedback. To help L2 writers alleviate some of the complexities in L2 essay writing, instructors can incorporate certain strategies and tools into the curriculum. Three such strategies and tools are scaffolding, process writing, and checklists.

2.5.4.1 Scaffolding

Scaffolding is a process or framework that assists learners with guidance or support when learning new skills or concepts. In the L2 writing context, scaffolding can involve a variety of techniques. Hyland (2019) explains scaffolding in L2 writing as a support mechanism that builds on the current knowledge or writing ability of the student, providing a sample or framework to build upon. Three commonly used scaffolding methods in L2 academic writing are mentor texts, schema building, and metacognition through feedback. Mentor texts are writing examples that include specific writing aspects that students strive to emulate, possibly reducing the amount of extraneous cognitive load. Kyun, Kalyuga, and Sweller (2013) conducted a study of mixed-leveled ELLs at a Korean university and found model essay use showed an improvement over increased writing practice. Additionally, the least proficient student writers found model use decreased difficulty in writing, which was not found in the proficient writers. This suggests that model essays can be especially beneficial in reducing difficulty (complexity) among less proficient L2 writers. Similar results have been found in other disciplines, such as mathematics and science (e.g., Kirschner, Sweller, & Clark, 2006).

Schema building is developing a plan or mental structure connected to previous knowledge or understanding. This can come in the form of a writing task on a subject that the students are familiar with or from prior knowledge of the vocabulary used in the task. They can more easily build new schemas from this familiar or previously acquired knowledge. By not having to produce a novel, unfamiliar topic, the complexity is reduced, and less cognitive load is likely to be needed (Sweller, 2011). This is illustrated in the difference between novice and skilled writers. Kalyuga (2010) makes the argument that according to cognitive load theory, the

difference between these two groups is due to the difference in knowledge base and not problem-solving strategies or better working memory. If the knowledge base on the specific topic or skills is large and well-connected, more cognitive resources may be freed to do deeper, more complex tasks (e.g., reflection).

Metacognition through self-feedback represents self-reflection, making evident what the learner does and does not understand. Metacognition through reflection from portfolio writing in an L2 context was investigated by Farahian, Avarzamani, and Rajabi (2021). They found that portfolio writing increased understanding and general reflection ability; however, they did not find an effect on critical reflection since the participants` writing skills were relatively low. They concluded reflection skills take time to improve, accounting for the lack of increased critical reflection. Moving from novice writer to skilled writer is an iterative process according to the knowledge-telling and knowledge-transforming models. This study further illustrates the need for the necessary schema to undertake tasks at a specific level, such as certain writing skills or being able to reflect critically. Vygotsky`s Zone of Proximal Development (ZPD) explains how scaffolding aids learning and enables the learner to take risks by bridging the gap between the learner`s current knowledge or ability and new learning (Richards & Rodgers, 2014; Williams et al., 2016). Metacognition through self-feedback differs from the other forms of scaffolding presently described in that it can also be a DD in addition to working as a form of scaffolding.

2.5.4.2 Process Writing

Process writing breaks down writing tasks into smaller units, alleviating complexity. While it varies in application, at its core, it narrows the writer`s focus at each stage. Graham

and Sandmel (2011) conducted a meta-analysis of 29 studies in the L1 context and found a modest improvement in writing quality among general education students. However, this benefit was not found among struggling writers, suggesting that the benefits may increase as writers` skills improve.

The research on cognitive load in process writing is limited, but it has been demonstrated that performing multiple functions during a given task can lead to cognitive overload (Lovell, 2020; Sweller, 2011). This is especially true when the content being studied is beyond the learner`s level. The split-attention effect, a phenomenon where the learner`s attention is divided by different task aspects or distractions, can occur (Lovell, 2020). Constant reflection and metacognition during the writing process may inhibit writing production in the L2 context if competing aspects require simultaneous attention. This can be from the continuous pulling away from writing content caused by thinking about mistakes in word choice, grammar structure, spelling, or other aspects that are particularly challenging to L2 writers. Generating and evaluating tasks can interrupt the writing process during the same task or writing session (Flower & Hayes, 1981), and writers are in a constant battle between performing and monitoring their cognitive processes (Hacker, Keener, & Kircher, 2009). Process writing looks to mitigate this issue.

L2 writing textbooks typically present a four-step writing process: pre-writing, drafting, revising, and editing. However, this linear structure may not be the most effective. The Flowers Paradigm (Flower and Hayes, 1981) merges drafting and revising into a single stage and with a planning stage for organizing and goal setting, while a review stage includes evaluating and revising, allowing writers to focus on specific aspects without distraction, reducing cognitive load. This writing process is implemented by being compartmentalized into four unique stages:

Madman, Architect, Carpenter, and Judge. The Madman stage involves unrestricted brainstorming. The Architect stage groups these ideas without ordering them. The Carpenter stage develops the first draft from these groups; however, the writer has the option to move on if stuck. The Judge stage refines the draft through editing and revising. The process is used to reduce distractions and unnecessary revisions. After the Judge stage completes the first draft, revisions can occur in any order, focusing on one stage at a time until the writing is satisfied. The Flowers Paradigm serves as a scaffold to build upon and manage the writing process, potentially reducing cognitive load by reducing the switching of attention, therefore limiting the intrinsic cognitive load burden. However, The Flowers Paradigm is an example of one form of process writing, and its effectiveness may vary across different groups or contexts.

3.0 OVERVIEW OF MODULES 1 & 2

3.1 Module 1

This chapter discusses the first two modules of this modular PhD. Module One laid the foundations for this by discussing key areas of DDs, CLT, and identifying generation as a DD that lacks strong evidence for or against its use in the L2 classroom. The first module provided a general historical background of SLA from the mid-1800s to its current state. It also provided an overview of DDs and studies on the efficacy of their use in learning. Module 1 further included an introduction to a counterargument for DDs, CLT. By reviewing the literature on DDs in the L2 classroom, it was clear that generation lacked in studies conducted compared to other DDs; moreover, studies in L2 contexts in the literature often conflicted with each other in their findings. Additionally, the dichotomy between the theory of DDs and the advantage of increasing difficulty with that of CLT, arguing that added complexity and difficulty lead to a cost in the encoding process and, therefore, learning, raised a clear need for research that is specific to the L2 context. Therefore, I decided to focus Module 2 on researching generation, a DD that has shown mixed results in the L2 context.

3.2 Module 2 Study Overview & Outline

Module 2 carried out a classroom action research study on the effects of generation on L2 vocabulary learning. This study consisted of two experiments investigating generation on

L2 learners` vocabulary retention ability. Vocabulary acquisition among L2 learners is core to communicating in the target language. According to Nation and Meara (2002), the lack of vocabulary acquisition in the target language can be the most important reason that keeps the learner from progressing. Therefore, this module investigated the effect that generation has on long-term vocabulary retention and looked to gain insight into how future studies may be adapted to research generation in areas outside of vocabulary learning.

3.2.1 Experiment 1

The initial experiment was conducted online synchronously via Zoom in the spring semester of the 2020-2021 academic year among 56 first-year Japanese university students at B1 to B1+ CEFR level (Japanese L1) studying at a Japanese university in an EFL context. The students were placed in estimated CEFR levels as a result of scores from the Computerized Assessment System for English Communication (CASEC). Due to COVID-19, the university did not conduct an in-person English placement test before classes started, and the CASEC was used in its place. This experiment studied three groups of students for six weeks. The generation task was the creation of novel sentences written in L2 (English) following the introduction of the vocabulary items through definitions, contextual reading, and comprehension activity. The experiment included three treatments, with 20 vocabulary items from each lesson (60 in total): use of definition only (baseline treatment), use of a bilingual

dictionary, and generation of sentences without a dictionary. The baseline condition received a definition in English and an example sentence for each vocabulary item at the beginning of each lesson. The definition and sample sentence were provided for each of the conditions. The dictionary condition was identical to the baseline, with the addition of participants being allowed to use a bilingual dictionary for reference during the lesson. The generation condition included the same definitions and sample sentences, but it also included a generation task. This generation task followed the reading comprehension questions from the lesson requiring students to generate an original sentence for each target vocabulary item. The sentence had to incorporate the vocabulary word in the proper context. All three groups were exposed to each treatment once by manipulating the conditions of the three lesson plans, and each treatment was tested three days and three weeks from the treatment. The initial three-day test was considered the baseline score for comparison with the three-week test. The long-term testing point for retention of target vocabulary was three weeks from the last exposure to the items. All tests were done through the Moodle quiz function with a four-option multiple-choice answer format for each question, and participants were informed only of practice tasks as warm-up activities and that they should not study for these. The contents of the tasks, which were actually post-tests, were not provided to the students to prevent studying outside of the controlled lesson and treatments. The study limited word knowledge to meaning recall ability of the provided definition from the corresponding lesson. The expected outcome was that the generation

treatment, causing increased difficulty during the learning process, would improve longer-term vocabulary retention. Table 3.1 shows the corresponding groups and treatments for each lesson.

Table 3.1: Week and Topic for Each Group, Treatment & Post-tests

Week (topic)	**Group A**	**Group B**	**Group C**
1 (Social Media) Treatment + 3-Day Test	Definition	Dictionary	Generation
2 (Enhanced Weathering) Treatment + 3-Day Test	Dictionary	Generation	Definition
3 (The Bachelor's Lesson) Treatment + 3-Day Test	Generation	Definition	Dictionary
4 (Social Media) 3-Week post-test	Definition post-test	Dictionary post-test	Generation post-test
5 (Enhanced Weathering) 3-Week post-test	Dictionary post-test	Generation post-test	Definition post-test
6 (The Bachelor's Lesson) 3-Week post-test	Generation post-test	Definition post-test	Dictionary post-test

Table 3.2 includes the mean and standard deviation (SD) scores for the three-day and three-week test results for all groups combined.

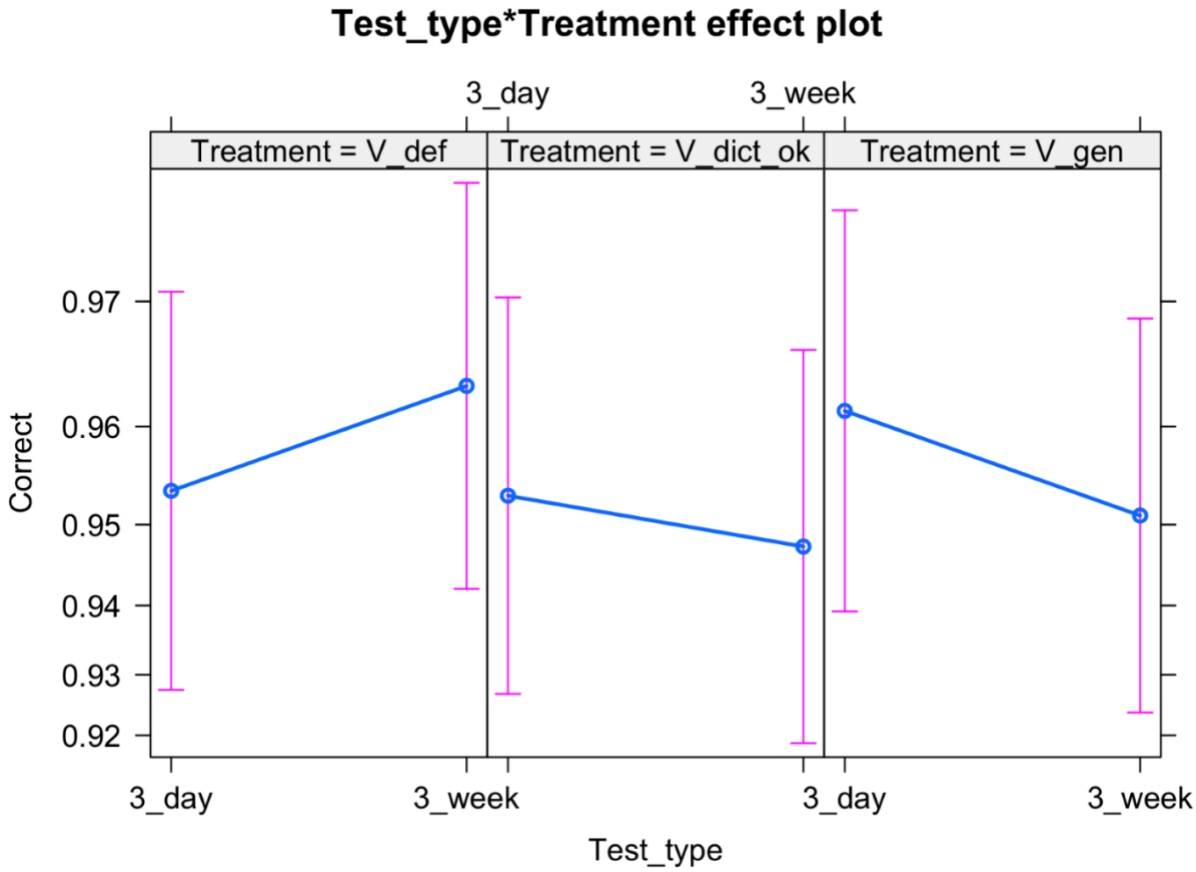
Table 3.2. Mean (with SD in brackets) for each treatment condition on the 3-day and 3-week tests. Scores are expressed as a proportion (percentage), where random guessing would equal 0.25

Condition	3-day test (baseline)	3-week test
Definition	0.91 (0.29)	0.92 (0.27)
Dictionary	0.91 (0.29)	0.90 (0.30)
Generation	0.91 (0.28)	0.90 (0.31)

The experiment used an initial model with fixed effects of condition, treatment coded so that definition was the reference condition, and three-day test type was the reference score. The model showed no effects from treatment condition or test type. However, there was a significant interaction of test type and condition: Generation ($\beta = -0.48$ $z = -2.15$, $p = .032$) and a marginal interaction of test type and condition: Dictionary ($\beta = -0.38$ $z = -1.69$, $p = .092$). Figure 3.1 shows a side-by-side comparison of treatment scores and test types, with three-week test type scores not showing a benefit for the generation treatment compared to the baseline treatment. Figure 3.1 further illustrates a lack of interaction between three-week scores for the dictionary treatment compared to the baseline. Covariates of faculty (major), group, and gender were analyzed to see if they improved the model. While neither group nor faculty improved the model, gender did show a significant improvement as a fixed effect ($\chi^2 = 9.91$, $p = .002$) and further improved the model as an interaction with treatment and test type ($\chi^2 = 20.12$, $p = .001$). However, a pairwise comparison showed no differences between treatments or test types for either gender. Additionally, there was an imbalance in gender (Females = 42, Males = 14), which makes the gender covariate results unreliable.

Figure 3.1: Comparison Plot of Model Treatments & Test Types

- V_def = Definition Treatment
- V_dict_ok = Dictionary Treatment
- V_gen = Generation Treatment



This experiment found no significant results, but issues with the design in how it was set up were apparent. Test scores were overall very high, which could indicate that the vocabulary items or the tests were not challenging enough for the participants. Another issue was the lack of a pre-test in addition to the three-day and three-week post-tests in order to provide a baseline for comparison. These issues were addressed in Experiment Two.

3.2.2 Experiment 2

Based on possible ceiling effects in Experiment One, the vocabulary difficulty level was adjusted for Experiment Two. In Experiment One, vocabulary items were estimated at the B2 level. Experiment Two used vocabulary estimated at the C1 level. Additionally, the amount of time allowed to answer test questions was shortened, and the multiple-choice answer tests were changed to a word bank list to increase the difficulty of the post-tests for better assessment. Furthermore, a pre-test was included before the start of the second experiment to establish a baseline for comparison with the three-day and three-week tests. To use words in the pre-test that were the same level (but not the exact same items to be tested in the treatments), the pre-test included 20 vocabulary items from a different unit in the textbook (Nation, 2009) from which the experiment lessons originated. The pre-test included the revised, more challenging testing structure of less time and vocabulary answers from a word bank instead of multiple-choice options. Table 3.3 shows the differences between the experiments.

Table 3.3: Differences between Experiments 1 and 2

Experiment 1	**Experiment 2**
15 minutes allowed for 3-day and 3-week tests	10 minutes allowed for tests
No pre-test	Pre-test of target similar vocabulary from the same source as vocabulary used in the experiment.
Vocabulary level of B2	Vocabulary level of C-1
Multiple-choice questions with 4 possible answers	Large list of answers (25) to choose from with extra choices

3.2.2.1 Methodology

Experiment Two was conducted on three groups of participants during the fall semester of 2020. The groups were arranged by the class the participants were enrolled in, and each class was one distinct group. As in Experiment One, the participants were L2 speakers of English at the B1 to B1+ CEFR level, with Japanese as their L1. The three groups had a range of 11-27 participants in each group at the start of the study. Data from 24 participants were excluded from the study due to incomplete tasks or quizzes. In total, 34 participants completed the experiment, comprising 25 females and 9 males. The participants' majors varied and did not include any students majoring in English language or literature. Participants were assigned to their respective English proficiency levels based on their CASEC test results prior to the study as described in Experiment One.

This experiment used lesson materials through reading and post-reading exercises with academic vocabulary lists and lessons at the C1 CEFR level from Nation (2009). The lessons were administered with different treatments to alter the study methods of the target vocabulary. Participants received color PDF copies of the lessons through Moodle without being provided with information on the origin of the lessons to prevent them from looking up the activities online. Each lesson comprised approximately four pages and followed a sequential order. The lessons included twenty target vocabulary words, accompanied by an English definition and an example sentence for each word. Additionally, the lessons incorporated a reading activity that utilized the target vocabulary in context and a post-reading comprehension activity. As in Experiment One, the lessons were administered with different treatments, namely Definition treatment, Dictionary treatment, and Generation treatment, to modify the study methods of the target vocabulary.

Experiment Two closely followed the methodology of Experiment One, with the modifications outlined in Table 3.3. The data analysis for Experiment Two was similar to that of the first experiment, with the addition of a vocabulary pre-test. This was implemented to account for the variability in vocabulary knowledge among participants within the same CEFR level and to establish a baseline for comparison. Results from the pre-test revealed a wider range of English proficiency levels among participants in each group than initially anticipated at the start of Experiment One. Pre-test scores ranged from 0 to 10 out of a possible 20 points (mean = 5.3, S.D. = 2.1).

Table 3.4 shows the corresponding groups and treatments for each lesson with the week next to the corresponding lesson title.

Table 3.4: Week and Topic for Each Group, Treatment & Post-tests

Week (topic)	**Group A**	**Group B**	**Group C**
0 Pre-test	Pre-test	Pre-test	Pre-test
1 (Greek Magical Papyri) Treatment + 3-Day test	Definition	Dictionary	Generation
2 (The Greedy Bee) Treatment + 3-Day test	Dictionary	Generation	Definition
3 (Epidemic in Zimbabwe) Treatment + 3-Day test	Generation	Definition	Dictionary
4 (Greek Magical Papyri)) 3-Week post-test	Definition post-test	Dictionary post-test	Generation post-test
5 (The Greedy Bee) 3-Week post-test	Dictionary post-test	Generation post-test	Definition post-test
6 (Epidemic in Zimbabwe) 3-Week post-test	Generation post-test	Definition post-test	Dictionary post-test

As in Experiment One, participants were allotted one hour to complete the initial vocabulary lesson and assigned treatment. This duration was sufficient for completing all treatments, including the Generation treatment, which required more time than the other treatments. Participants assigned to the Definition and Dictionary treatments were permitted to review and continue studying until the end of the allotted hour to compensate for the additional time spent on vocabulary in the Generation treatment.

The data from Experiment Two were analyzed in the same way as that employed in Experiment One, with the inclusion of a pre-test to ascertain the novelty of the vocabulary to the participants. As with Experiment One, the baseline score was established using the results of the three-day test, which were subsequently compared to those obtained in the three-week test. The analysis focused on examining the impact of the treatments (Definition, Dictionary, and Generation) and the test type (three-day and three-week).

3.2.2.2 Results in Aggregate for Experiment 2

The aggregate results are provided in Table 3.5 in percentages correct with standard deviation.

Table 3.5: Mean Scores (standard deviation) for Experiment 2 in Percentages

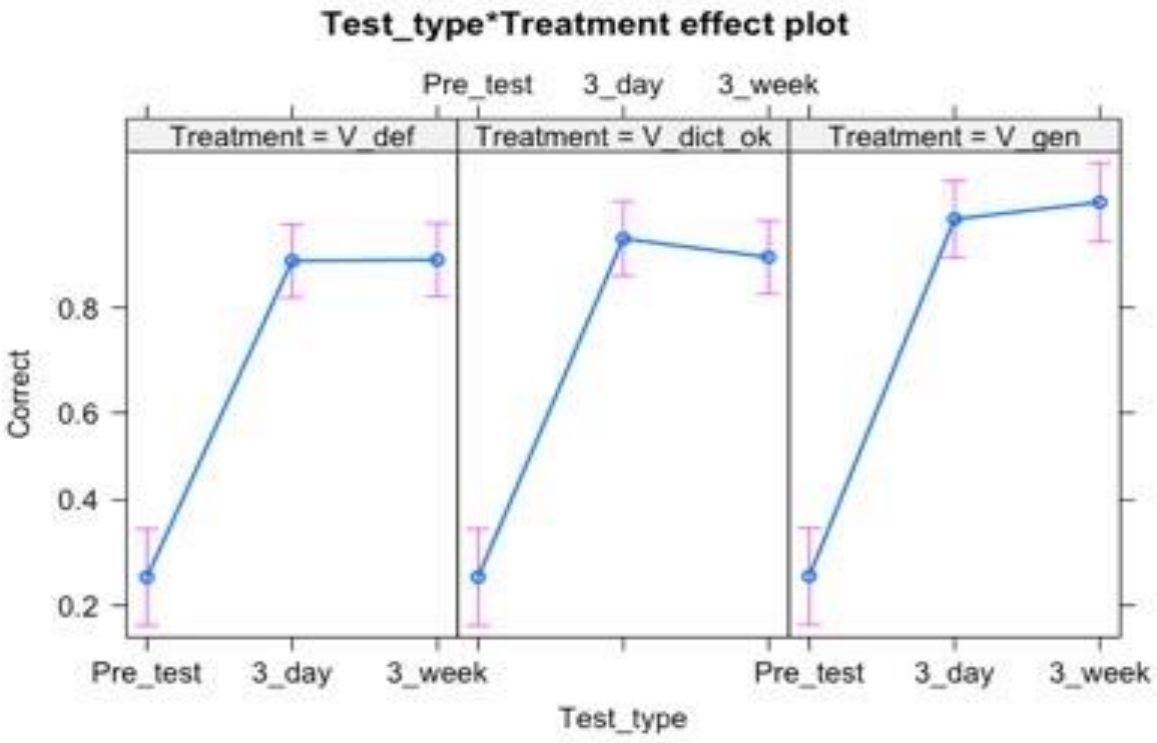
Condition	**Pre-test***	**3-day test**	**3-week test**
Definition	.263 (.441)	.816 (.388)	.819 (.385)
Dictionary	.263 (.441)	.838 (.369)	.818 (.386)
Generation	.263 (.441)	.869 (.338)	.885 (.319)

*The pre-test used different vocabulary items taken from the same level as those used in the post-tests. The mean and standard deviation scores shown in percentages reflect the overall average for the group, as students had not been subdivided into conditions at this point.

An initial model was constructed with fixed effects of condition (with Definition as the reference level) and test type (with three-day as the reference level). To account for differences in baseline vocabulary knowledge, the pre-test score for each participant was included. Additionally, Experiment Two examined the covariates group, faculty (major), and gender to determine their effect on the treatments.

Figure 3.2: Comparison Plot of Treatments & Test Types

V_gen = Generation Treatment
V_def = Definition Treatment
V_dict_ok = Dictionary Treatment



As visualized in Figure 3.2, the final model revealed a marginal effect of the Generation condition ($\beta = 0.44$, $z = 1.75$, $p = .080$). However, no other main effects or interactions were significant (all $zs < 2.00$, all $ps > .05$). A pairwise analysis was conducted using the emmeans

package with a Bonferroni correction for multiple comparisons. The results showed that after three weeks, scores in the Generation condition were not statistically significantly higher than those in the Dictionary condition ($z = 3.21$, $p = 0.20$) or than those in the Definition condition ($z = 2.76$, $p = .089$). No other comparisons were significantly different

3.2.2.3 Covariates: Participant Faculty, Group, & Gender

Experiment Two included faculty, group, and gender as covariates, and none of these were significant.

3.2.2.4 Results Summary

The results of Experiment Two did not reveal any significant differences according to the type of treatment. Furthermore, the inclusion of covariates (group, faculty, gender) did not result in any improvement to the model. Adding a pre-test provided confidence that the vocabulary level was sufficiently difficult, which was not the case in Experiment One, and the changes implemented in Experiment Two, such as reduced time allowed for post-tests and the use of a large bank of possible answers, provided greater confidence in the results. Overall, no clear effects of generation were observed in either of the studies conducted in Module 2.

3.2.3 Conclusion

Module 2 explored the use of generation on L2 vocabulary by performing two L2 vocabulary experiments on Japanese university ELLs. Issues around methodology (difficulty level of the post-tests and vocabulary, and the lack of a pre-test) in Experiment One were corrected in Experiment Two, which led to more confidence in the findings. Additionally, due

to the Covid19 pandemic, Module 2 experiments were conducted online, which differed from relevant studies in the literature and affected the ability to ensure adherence to the instructions provided. Nevertheless, the findings provided more evidence that L2 vocabulary learning through generation may not provide a substantial benefit to L2 vocabulary learning. To better address this question, Module 3 includes a study of generation and L2 vocabulary learning by expanding on word knowledge beyond definition recall alone. Furthermore, the level of confidence in knowledge by the learners is explored to look at vocabulary knowledge from different angles. Additionally, the effect of generation on L2 learning in reading comprehension and writing will be investigated to see if generation`s efficacy findings are specific to L2 vocabulary learning or to the L2 context in general.

4.0 Study 1 Effects of Generation on L2 Vocabulary Learning

4.1 General Introduction and Rationale

This study examined the effect of generation on both long-term retention of L2 vocabulary and the confidence that participants have in their knowledge of the target vocabulary. While the effect of spacing, retrieval, and interleaving DDs have been shown to be effective in L2 vocabulary learning, the effect of generation on L2 vocabulary acquisition is not as clear; therefore, this study focuses specifically on generation, with the goal of gaining insight into the efficacy of L2 vocabulary meaning recall over two periods (three and 15 weeks) from initial exposure in the lesson. This was a classroom action research study that began during the first semester of the academic year in Japan and continued early in the second semester after the university summer break. Participants were first-year ELLs in a Japanese university, and all were Japanese L1 speakers studying in an EFL context.

This study investigated whether the recognized positive effects from generation tasks on long-term learning (e.g., Jacoby & Chestnut, 1978) where learning is done in L1 tasks such as summarization (e.g., Wittrock, 1974b; 1989), reading tasks (e.g., DeWinstanley & Bjork, 2004), and through L1 word lists (e.g., McDaniel, Waddill, & Einstein, 1988) are similarly effective among ELLs. Alternatively, ELLs may encounter higher demands on cognitive load and, therefore, may not show the same level of benefit from generation or may even show a cost (e.g., Barcroft, 2007; 2009; 2015).

Many studies limit their measure of "word knowledge" to recall the definition provided at the time when the target vocabulary item was first introduced. However, in this study, participants were required to demonstrate their knowledge by producing their own definitions

and context-based examples of use illustrating meaning recall. By doing this, the chance of guessing the meaning from word forms such as prefixes or suffixes was reduced since answers were not provided in the form of a word bank or multiple-choice answers, which could be used to increase the chances of guessing correctly.

In addition to testing knowledge, this study also investigated the confidence that learners have about each learned item. Participants' confidence in their vocabulary knowledge may not match their actual knowledge, and in some cases, learners may over- (or under-) estimate the knowledge that they actually have. This false fluency effect may occur when knowledge comes from similarly spelled words or meanings used in different contexts. The confidence in using a word may, therefore, be limited to a specific form or context. This could lead to a false impression of comprehension when the learner lacks the contextual knowledge required to understand the target word in a specific situation. Therefore, this study assessed the participants' confidence in each target word and compared this confidence to the actual ability to use or explain the term correctly in context. Confidence ratings have been shown by Pellicer-Sanchez et al (2021) to show a relationship between accuracy and confidence in both L1 and L2 learners among word collocation forms.

The vocabulary items were selected from the list in the lessons used in Nation (2018). A pre-test of the same vocabulary in the post-tests was given at the onset of the study. According to Webb (2020), assessment of vocabulary knowledge (pre-and post) is a key aspect of both classroom and research studies. The stages of vocabulary assessment should encompass pre-testing, which serves to measure existing knowledge or prepare learners for forthcoming activities, as well as short-term and longer-term post-testing to ascertain the retention of target vocabulary by learners. The purpose of the pre-test in the present study (in both Confidence and

Knowledge scores) was to establish a baseline for comparison with subsequent tests. This test served two purposes. The first was to establish a self-reported participant confidence score, identified as "Confidence" in this study, indicating the confidence they had in using each vocabulary item. Next to each target word, participants were required to indicate their knowledge by scoring the word as 1 (do not know the word), 2 (seen before but do not know the meaning), 3 (understand the meaning if read or heard in a sentence but cannot use it), or 4 (understand the meaning and can confidently use it in speaking and writing). A four-point scale used is similar to that used in other vocabulary confidence studies (e.g., Pellicer-Sanchez et al., 2021).

The second purpose of the pre-test was to establish whether students truly understood the meaning and could use the vocabulary items in the form provided. Following each self-reported score, participants were asked to write the meaning and give an example usage to demonstrate how well they actually understood the meaning. This was to prevent false fluency situations and to establish a baseline for comparison of post-tests. This is referred to as "Knowledge score" (for whether the participant actually knew the meaning of the item) in this study. Knowledge score answers were marked either correct or incorrect. The example and definition needed to clearly illustrate knowledge of the word and its usage to be considered correct. Since the lesson had not yet been presented to the students during the pre-test, correct use in any context was accepted. Correctly using a word in the pre-test in a different context from the lesson`s context was not evident when scoring the pre-test. This changed for future post-tests once the vocabulary lesson with the specific context was learned, and all answers needed to show recall of the word`s meaning of the context from the lesson. Partially correct scores were not included in this study. When scoring the completed pre-and post-tests, many

participants included vague answers, which were likely guesses that could be interpreted as possible understanding. A simplified correct or incorrect determination was used to better determine if there was an improvement in knowledge as a result of the treatment and to avoid the messiness of qualifying partially correct answers.

The study did not include follow-up or additional activities beyond the original introduction of the vocabulary and lesson activities. However, while participants were instructed not to do any independent studying of the terms introduced, verifying that this did not occur was impossible. To minimize the chances of participants studying through outside sources or from other activities not related to the experiment or target vocabulary, all tasks were conducted in class on paper. The instructor collected lesson materials before the participants left the classroom.

Data from pre- and post-tests were collected over a 15-week period during face-to-face classes. Because the tests were conducted in view of the instructor, compliance could be monitored. The rationale behind the time distribution of the three-week test is grounded in Nation`s (2008) purpose and features of vocabulary tests and is further supported by Schmitt (2010). Nation (2008) and Schmitt (2010) both argue that post-tests of a minimum of three weeks are optimal to assess whether vocabulary has been durably acquired. The second Post-test (Post-test 2) was given 15 weeks following the initial exposure to the target vocabulary and the lesson, which allowed a much longer period between exposures than is often reported in the literature, such as Barcroft (2007) comparing two days and one week or Sun and Fang (2022) comparing one, two, and three weeks. Furthermore, students did not have access to the lessons, words, or post-test 1 as the instructor collected all the materials.

The effect on longer-term meaning recall could be evaluated by assessing the effects at three weeks from the lesson and another assessment at 15 weeks. This longer-term learning is more in line with the actual use of L2 vocabulary outside of in-class quizzes, where vocabulary items are often forgotten soon after taking a quiz with little to show for the effort in learning the meaning of the target words. Many university classes are only a single semester in length (12-14 weeks), which limits the amount of time that teachers and researchers can assess recall. The 15-week post-tests were used to investigate the effect of generation following a summer vacation where participants may have been less likely to partake in English language learning. Furthermore, the extended 12-week period beyond the first three-week post-test (15 weeks from initial vocabulary exposure and the lesson) was needed due to the eight-week summer break between the first and second semesters.

4.2 Methodology

The methodology of this study was based on a predetermined curriculum and participants. As this was a classroom action research project, the study took place in a classroom setting with university-assigned students as participants. Therefore, the methodology was restricted in options and based on convenience, which accounts for the limited number of participants and the timeframe of the post-tests.

4.2.1 Participants

The participants in this study consisted of 38 first-year university students in Japan (L1 Japanese). They were, on average, 18 years of age, which is standard in Japan for first-year university students. The participants were L2 speakers of English at the B1 to B1+ Common

European Framework (CEFR) level. Prior to the beginning of the school year, the participants took an English level placement test (Test of English for Academic Purposes, TEAP) and were placed in a class that corresponded to B1 to B1+ CEFR level. The TEAP is used by the university where the study was conducted to place all incoming first-year students into English classes according to their estimated English proficiency levels. The test assessed reading and listening skills, but the version of the TEAP used by the university did not include writing or speaking assessments. The TEAP results identifying the participants at the B1 to B1+ level were also used to estimate the students` L2 vocabulary level. The majors of the participants varied (e.g., global studies, journalism) but did not include majors where further English studying was required during the school year when the study was conducted. None of the participants majored in English or English literature, which could have led to further exposure to the target vocabulary in other classes outside of the controlled study.

This vocabulary study consisted of two groups with a total of 38 participants from both groups at the beginning of the study (*Group A = 16, Group B = 22*). Groups were the same as university classes in which the students were enrolled (Class A = Group A, Class B = Group B). Data from 15 participants (*Group A = 8, Group B = 7*) were removed prior to analysis due to incomplete data. This was primarily a result of participants missing multiple weeks of class from Covid-19 infection and, therefore, they were unable to complete the lessons or post-tests in person. The final data analyzed were from 23 participants, with 6 males and 17 females (Group A males = 2, females = 6; Group B males = 4, females = 11) who completed all aspects of the study.

Both groups were part of classes made up of students from first-year compulsory Academic Communication classes meeting bi-weekly for 14 weeks in the first semester and 14

weeks in the second semester in face-to-face, on-campus classes. Students from semester one were carried over to the second semester, allowing for longer-term assessment. The classes for both groups were conducted on Mondays and Thursdays, with Group A meeting from 9:00 to 10:40 and Group B meeting from 15:25 to 17:05.

This study did not provide compensation, nor were the test scores counted toward the course grade. This was explained to the participants before the start of the study and reiterated throughout. While the test scores were not included in the final grade, participation in the lessons was considered a compulsory aspect of the class and affected the participation part of the final grade. Participants were provided a consent form clearly detailing that they could opt-out at any time during the study from their data being used in the analysis by contacting the instructor (see Appendix 1a). This was approved by the University of Birmingham Research and Ethics Committee.

4.2.2 Study Materials

The study materials included an initial pre-test (Appendices 2a & 2b), paper-based color copies of the lessons consisting of approximately 7-8 pages. The contents of the lessons were included in sequential order: 20 target vocabulary items with definitions, images, and sample sentences, a reading activity of approximately 270 words in length with all target vocabulary in context, and a post-reading comprehension activity. These initial parts of the lesson were the same for both treatments. The lessons were derived from a publicly available ELL textbook (Nation, 2018), estimated to be at the B2 CEFR level. This study followed *The History of Chocolate* and *How the Dinosaurs Really Died* themes from Nation (2018).

4.2.3 Rationale for Target Vocabulary Chosen

Since the participants were at the B1 to B1+ CEFR level, finding vocabulary that would be challenging but not overly difficult was needed. Additionally, it was important to find words that students were less likely to be exposed to in daily life (e.g., on television or social media). The chosen ELL textbook (Nation, 2018) included pre-selected vocabulary lists for each unit at the B2-level and tasks that allowed for a more challenging level of text and vocabulary, increasing the chances of the target items being unknown. B2-level vocabulary should also be comprehensible after being introduced, as it is only one CEFR level above the estimated English level of the participants. In addition to the vocabulary items being B2 level, they were also all derived from Coxhead's (2000) Academic Word List (AWL). Therefore, the vocabulary items incorporated two aspects of difficulty: vocabulary level above the participants` class level and the use of academic vocabulary items which participants are less likely to be exposed to in daily life outside of school. The 20 vocabulary words used in each lesson (40 in total between the two lessons) allowed for enough variety to ensure a balanced difficulty level between treatments, even if some words were easier than others. Target vocabulary items did not account for interlexical or intralexical factors.

Building upon the base lesson in the textbook, treatments were administered to alter the methods the participants used in learning the target vocabulary. The details on how treatments differed and the implementation are provided in the Procedure Section (4.1.4). The lessons were provided to each participant immediately after completing and submitting the pre-test (see Appendix 3 for lessons). To limit the possibility of students identifying and independently obtaining a copy of the textbook used as the base for the lesson, all identification to the author and textbook were deleted from the materials provided to the participants.

4.2.4 Procedure

This study was conducted over a 15-week period, and both groups completed the same lessons and saw the same conditions and the same set of vocabulary items. The groups were not counterbalanced and were identified by separate groups to account for the difference in possible class dynamics and time (class period) when the lessons took place. However, due to Group A only having data from eight participants at the end of the study, group was not included in the final analysis as a covariate.[1] The baseline and generation vocabulary was the same for both groups. In each lesson, students saw half of the items in each of the two conditions. Baseline condition vocabulary items (10 from the 20-item list) were chosen from previously prepared activities in the textbook. The remaining 10 items not used in the baseline task were used as the generation items. The study administered two treatments: a baseline control condition where vocabulary tasks included non-generative tasks (e.g., vocabulary to definition matching) and a generation task where participants were required to produce novel sentences using the target vocabulary in context to clearly show the use and meaning of the term as given in the lesson. Participants were allowed to use a bilingual dictionary, but the example sentences provided in the dictionary were not permissible as novel sentences. The novel sentences needed to be personalized to show originality and understanding.

[1] Many participants contracted Covid 19 during the study and were unable to complete the tasks in their assigned Group, which affected the Group sizes. Since Group A only had data from eight participants at the end of the study to analyze, Group was not included in the final analysis.

Steps taken by participants in order were:

- A pre-test of 20 target vocabulary items from the first lesson (*The History of Chocolate*) was completed by all participants (Confidence and Knowledge scores collected for all 20 items).
- The first part of Lesson One was completed (including the introduction of 20 target vocabulary items, definitions, and sample sentences at the beginning of the lesson before starting the reading task comprehension questions.
- The second part of Lesson One included the Generation task for ten target items and baseline tasks for the remaining ten items.

In week two, participants completed the second pre-test and Lesson Two identically as was done for the first pre-test and lesson. The week two pre-test included the 20 target vocabulary items before introducing the words in the lesson (first part), and then applied the baseline or generation conditions to ten words each (second part).

Specific instructions were given before each activity (pre-test, lesson, and post-tests) to ensure consistency. In week one of the study, *The History of Chocolate* lesson was completed by Groups A and B on the same day but during different class periods. Participants spent one 100-minute class period completing the pre-test and lesson, which was more than enough time to complete the lesson and treatment tasks. All participants practiced half the vocabulary items in the baseline condition and half in the generation condition. The words used in each condition were the same for all participants and groups.

The baseline activities included the following: fill-in tasks with words from a word bank, synonym comparison task, definition matching task in the form of definition-to-vocabulary multiple-choice questions, and a crossword puzzle task. The instructor developed the extra crossword puzzle task to compensate for the extended time needed to complete the generation task. Consequently, there was an equal length of exposure to the vocabulary items in each treatment. Both sets of activities took approximately 40 minutes to complete. This is separate from the time needed for the reading and comprehension activities, which were the same for both vocabulary conditions. Baseline tasks only exposed the participants to the 10 baseline words and did not provide further exposure to the 10 vocabulary items used in the generation task (and vice versa).

The generation task in the lesson had a list of vocabulary items with space for the students to handwrite two to three novel sentences using the term in a way that related to them personally. An example of a term not part of the lesson plan was provided at the beginning of this task of what was expected of the participants. The specific model used in both lessons is provided in Figure 4.1. The participants produced a similar personalized example for all ten generation condition vocabulary items.

Figure 4.1: Model answer at the Top of the Generation Lesson Task

For example:
procrastination*: It means not doing something immediately even though it should be done right away.*

Procrastination is a problem for me at school. Last week I waited until just before the deadline for my English essay before starting it. I was up all night writing it, and I did not get a good grade because I **procrastinated.**

Once participants completed the lesson on their own, the instructor visually verified that participants had properly completed the activities for the baseline activities and the generation treatment. For the generation sentences, the focus was on ensuring that sentences were adequate in content and correct usage during and before students submitted lesson contents at the end of the class. In the baseline activities, the focus was on verifying that students completed the questions using the target baseline vocabulary and did not just guess to fill in answers. Once the instructor verified that the tasks were properly completed, the participants were put into groups of three to compare their answers to the baseline tasks and to read their generation treatment sentences to their group. All correct answers to the baseline tasks and reading comprehension questions were provided to each group of three participants to check once they finished sharing their answers in groups.

The instructor collected lesson materials at the end of each class period, and no in-class or homework exposure was provided to the vocabulary items following the lesson treatments. Three weeks following the completion of the lesson plan, post-test 1 was conducted. All participants were provided with the same test as the pre-test (Confidence scores and Knowledge scores). The second lesson, *How the Dinosaurs Really Died*, was conducted

identically as the first lesson but one week later from the first lesson with the follow-up post-test 1 three weeks after the completion of the lesson.

The two post-tests 2s (one per topic) were held 15 weeks after the initial introduction of the vocabulary items. Hence, the post-tests took place one week apart. This ensured that each post-test 2 would have the same number of weeks between the lesson and post-test 2 (15 weeks from the lesson and 12 weeks from post-test 1). This followed the summer vacation, and participants were surprised by the second post-test (for the first lesson), which suggested that they had not been actively thinking about the vocabulary and lesson during the 12-week period following the first post-test. Post-test 2 was identical to the pre-test and post-test 1; therefore, students were asked for Confidence and Knowledge ratings for each word on the test. Table 4.1 below shows the corresponding groups and treatments for each lesson with pre- and post-tests.

Table 4.1: Week and Topic for Each Group, Treatment & Post-tests

Week (topic)	**Groups A & B** (Vocabulary conditions regarding treatment and specific words)
Week 1 (The History of Chocolate) Pre-test + Treatment	10 vocabulary items -Generation treatment 10 vocabulary items -Baseline treatment 20 vocabulary items in total
Week 2 (How Did the Dinosaurs Really Die?) Pre-test + Treatment	10 vocabulary items -Generation treatment 10 vocabulary items -Baseline treatment 20 vocabulary items in total
Week 4 (The History of Chocolate)	Post-test 1 (20 vocabulary items per post-test)
Week 5 (How Did the Dinosaurs Really Die?)	Post-test 1 (20 vocabulary items per post-test)
Week 15 (The History of Chocolate) Post-test 2	Post-test 2 (20 vocabulary items per post-test)
Week 15 (How Did the Dinosaurs Really Die?) Post-test 2	Post-test 2 (20 vocabulary items per post-test)

Participants were given thirty minutes to complete the pre and post-tests, which was enough time for everyone to complete all parts of the tests. While all pre and post-tests were conducted in class in front of the instructor, participants were instructed not to look at any materials while taking the tests and to answer from memory to the best of their ability without fear of their results affecting the class grade. This was reiterated before each test and throughout the study. As part of the tests included self-reported Confidence scores on vocabulary knowledge, it was especially important that participants understood that if they reported a lower score, it would not be counted against their grade.

4.2.5 Data Collection and Assessment

All pre and post-tests were identical in question order to avoid the possibility that results might be affected by variation between tests. The use of identical formatting of pre- and post-tests is suggested by Nation and Webb (2011) as optimal to ensure that any differences in results are not caused by a difference in difficulty between the tests. The tests were completely in English without any Japanese in the directions. For self-reported Confidence scores, a space was provided for the participant to write the number corresponding to their self-reported knowledge of the word adjacent to each test item. At the top of all tests was a reference to the Confidence score numbers and corresponding references (e.g., 3, I understand this word when I see or hear it in a sentence, but I don`t know how to use it.).

The Knowledge answers had space for multiple sentences if needed to describe the meaning or provide a sample of the word in context. For Knowledge scores, the assessment of correct or incorrect was determined by the instructor. Spelling and grammar mistakes were allowed if they did not take away from clearly showing the participant understood the meaning and ability to show the correct usage. When answers were vague or unclear, the vocabulary term was considered incorrectly answered. Likewise, when a meaning of the word that was different from what was taught in the lesson was provided, this was considered incorrect.

4.2.6 Results

To compare the effects of the two treatment conditions, two different linear models were constructed using Jamovi Project version 2.3, R version 4.1, and GAMLj (Gallucci, 2019). A linear mixed model was used for the Confidence scores (1-4) for each vocabulary

item. This kind of data can be effectively analyzed using a mixed-effects model because it allows for participants and specific items (e.g., individual words) to be treated as random variables, thus it can account for intra-participant effects in addition to inter-group results. For the Knowledge scores, a generalized linear model with binomial distribution was used as the scores were marked either correct or incorrect.

All data from the two groups were compiled as one file for analysis. For both Confidence and Knowledge scores, each vocabulary item was used as one data point per item, with individual Confidence and Knowledge scores at each testing point. Data from the self-reported Confidence scores were analyzed by compiling the scores for each vocabulary question in the Pre and Post-tests. The possible Confidence scores included a "1" for no knowledge, "2" for having seen the word before, but do not know the meaning, "3" for understanding the meaning, but cannot use it in context, and "4" for confidence in the meaning and ability to use. For the Knowledge scores, test questions were assigned a "1" for correct and a "0" for incorrect. This allowed for 20 data points per test, for each participant to have 60 data points for the pre-test and the two Post-tests for each of Confidence and Knowledge (60 x two treatments = 120 for each lesson). The two lesson pre- and post-tests combined had 240 data points (*The History of Chocolate* = 20 data points x pre-test and two post-tests = 120 + *How Did the Dinosaurs Really Die?* = 20 data points x pre-test and two post-tests = 120). 23 participants completed the study, which gave an overall total of 2,760 data points (23 participants x 120 post-test data points) for each of the two test types (Confidence and Knowledge).

4.2.7 Results in Aggregate for Confidence Scores

The overall aggregate results of participants' self-reported Confidence scores are shown in Table 4.2. The scores range from a low of "1" (no knowledge of the vocabulary item) to a maximum of "4" (confident in understanding and ability to use the vocabulary).

Table 4.2: Mean Confidence Scores by Treatment (and Test Type with standard deviation (SD) and 95% Confidence Intervals (CIs)

	Baseline condition			**Generation condition**		
	Pre-test	Post-test 1	Post-test 2	Pre-test	Post-test 1	Post-test 2
Mean	3.08	3.43	3.41	3.09	3.55	3.47
SD	1.002	0.859	0.755	0.961	0.772	0.733
95% CIs	2.99, 3.17	3.36, 3.51	3.34, 3.48	3.00, 3.17	3.48, 3.62	3.41, 3.54

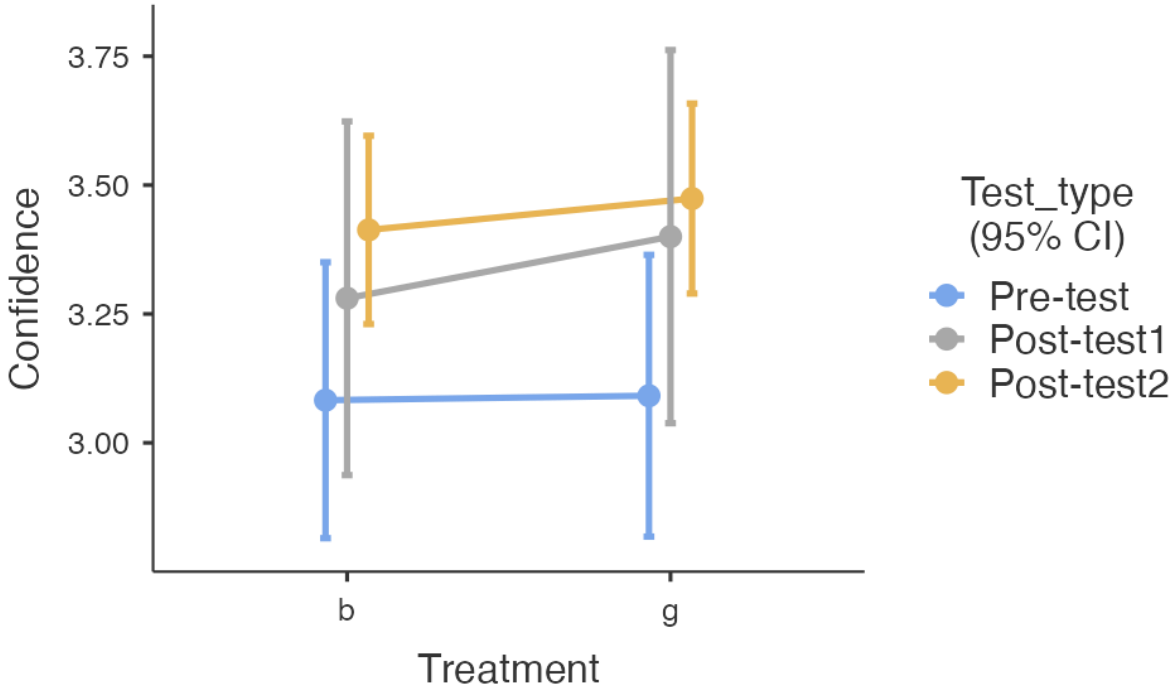
An initial model included fixed effects of condition, treatment coded with baseline as the reference level, and test type, treatment coded with pre-test as the reference level. Included were random intercepts for subject and vocabulary items, along with by-subject slopes for condition, and by-item random slopes for the effects of test type (pre and post-tests). Slopes for condition by subject were used since the subjects were exposed to variables differently in this study, and both control and generation items were seen by all subjects. This was included to maximize the random effect structure and, therefore, the generalizability of the model (Barr, Levy, Scheepers, & Tily, 2013). The model investigated self-reported Confidence scores for all three tests (pre-test, post1, post2). The overall effect of the treatment was found not to be

significant (β = 0.06, t = 0.50, p = .622). There were simple effects of post-test 2 (t = 6.29, p < .001), but not post-test 1 (t = 1.68, p = .106), showing that the later post-test (but not the first) showed a significant improvement over the pre-test scores. Interactions of treatment with post-test 1 (t = 1.50, p = .134) and post-test 2 (t = 0.71, p = .481) did not show a significant effect (see Table 4.3). The GAMLj package was used to conduct Post Hoc test pairwise analysis, with Bonferroni correction for multiple comparisons, which confirmed that scores in the generation condition were not significantly higher at the pre-test (z = 0.34, p = 1 .00), at post-test 1 (z = -2.62, p = .132) or at post-test 2 (z = -1.64, p = 1.00). Figure 4.2 compares the treatments at each test point.

Table 4.3: Treatment Interactions t and p-values with Confidence Intervals Included

				95% Confidence Interval				
Names	**Effect**	**Estimate**	**SE**	**Lower**	**Upper**	**df**	**t**	**p**
(Intercept)	(Intercept)	3.2902	0.0912	3.1116	3.469	51.8	36.094	< .001
Treatment1	g - b	0.0630	0.1277	-0.1872	0.313	38.8	0.494	0.624
Test_type1	Post-test1 - Pre-test	0.2533	0.1508	-0.0423	0.549	24.4	1.680	0.106
Test_type2	Post-test2 - Pre-test	0.3565	0.0569	0.2451	0.468	34.5	6.270	< .001
Treatment1 ✻ Test_type1	g - b ✻ Post-test1 - Pre-test	0.1109	0.0969	-0.0790	0.301	42.0	1.145	0.259
Treatment1 ✻ Test_type2	g - b ✻ Post-test2 - Pre-test	0.0522	0.1001	-0.1440	0.248	40.5	0.521	0.605

Figure 4.2: Side by Side Comparison of Confidence ratings by Treatment (B = Baseline, G = Generation) including 95% Confidence Intervals (CI)



4.2.8 Results in Aggregate for Knowledge Scores

The overall results of Knowledge scores are shown in Table 4.4, illustrating the percentage of vocabulary items correctly answered.

Table 4.4: Mean Knowledge Scores for Items Answered Correctly by Treatment, Test Type with standard deviation (SD) and 95% Confidence Intervals (CIs)

Baseline condition				**Generation condition**		
Tests	Pre	Post-1	Post-2	Pre	Post-1	Post-2
Mean	0.40	0.52	0.46	0.36	0.59	0.51
SD	0.490	0.500	0.499	0.479	0.492	0.500
95%, Cis	0.353, 0.443	0.570, 0.500	0.415, 0.507	0.313, 0.400	0.548, 0.639	0.467, 0.559

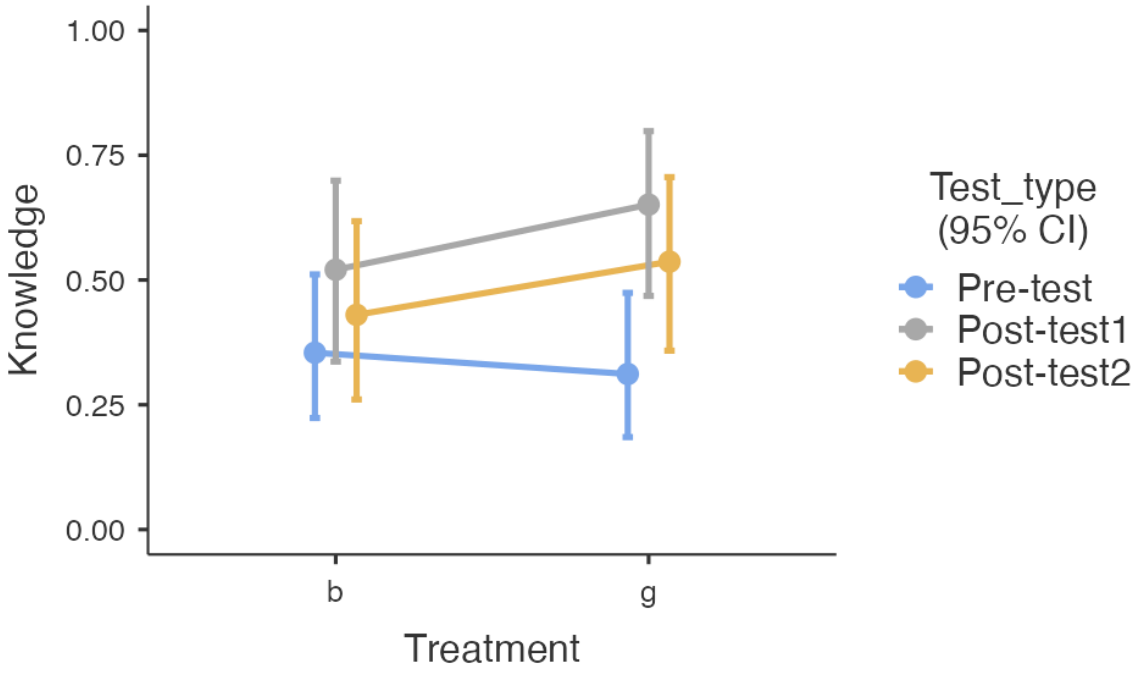
Note. The CI of the mean assumes sample means follow a t-distribution with N - 1 degrees of freedom.

A generalized mixed model was fitted with treatment-coded fixed effects of treatment (reference = Baseline condition) and test type (reference = pre-test). Random intercepts were included along with by-subject slopes for condition and test type, and by-item slopes for test type since one item was never seen in different conditions. This was included to maximize the random effect structure and, therefore the generalizability of the model (Barr, Levy, Scheepers, & Tily, 2013). The overall effects of treatment (β = -0.19, z = - 0.44, p = .661) were found not to be significant. There were simple effects of post-test 1 (z = 3.95, p = <.001), but not post-test 2 (z = 1.61, p = .108), showing that the former post-test (but not the second) showed a significant improvement over the pre-test scores. Interactions of treatment with post-test 1 (z = 2.62, p = .009) and Post-test 2 (z = 2.56, p = .012) showed a significant effect (see Table 4.5). Interactions show an overall increase in scores from the pre-test for generation treatment over the control. However, Post Hoc pairwise analyses with Bonferroni correction for multiple comparisons showed there were no significant effects of treatment at each test point: pre-test (z = 0.44, p = 1.00), post-test 1 (*z* = -1.85, p = .972), or post-test 2 (z = -1.03, p = 1.00) as shown in Figure 4.3.

Table 4.5: Treatment Interactions z and p-values with Confidence Intervals Included

					95% Exp(B) Confidence Interval			
Names	**Effect**	**Estimate**	**SE**	**exp(B)**	**Lower**	**Upper**	**z**	**p**
(Intercept)	(Intercept)	-0.600	0.329	0.549	0.288	1.05	-1.824	0.068
Treatment1	g - b	-0.192	0.438	0.825	0.350	1.95	-0.438	0.661
Test_type1	Post-test1 - Pre-test	0.682	0.173	1.979	1.411	2.78	3.954	< .001
Test_type2	Post-test2 - Pre-test	0.319	0.198	1.376	0.933	2.03	1.609	0.108
Treatment1 ✻ Test_type1	g - b ✻ Post-test1 - Pre-test	0.734	0.280	2.083	1.204	3.60	2.624	0.009
Treatment1 ✻ Test_type2	g - b ✻ Post-test2 - Pre-test	0.620	0.247	1.860	1.147	3.02	2.515	0.012

Figure 4.3: Interaction of Treatment (b = Baseline, g = Generation) and Test type for Knowledge scores with 95% Confidence Intervals



4.3 Discussion

In this study, data was collected on perceived knowledge (Confidence scores) and actual knowledge (Knowledge scores) based on newly learned vocabulary. Data obtained for both measures included pre-test, post-test 1, and post-test 2 scores. Neither measure showed a significant benefit from generation, compared to the baseline activities, and generation also did not show a negative outcome (e.g., no cost compared to baseline activities). While the findings were not statistically significant regarding simple comparisons as borne out by the Post Hoc tests at each test point, interactions in the Knowledge scores, as shown in Table 4.5 and Figure 4.3, imply a benefit for generation conditions from pre-test to post-tests. Generation treatment

for both Confidence and Knowledge scores did not negatively affect recall. Barcroft (2009) found generation in L2 vocabulary recall to be ineffective when used as the primary DD. The findings in this present study likewise do not show a clear overall benefit, although the significant interactions of condition and test time in Knowledge scores do suggest that the relative increase in the generation condition may have been more than in the control condition. Unlike in that of Barcroft (2009), there was no negative effect from generation. Furthermore, there were key differences between Barcroft`s study and the present study. One was that Barcroft looked at synonym generation rather than the generation of novel sentences. Synonyms may be less likely to have a personal connection to the learner than novel sentences. Another difference was in the population studied; Barcroft looked at Spanish L1 speakers learning English as L2. The Spanish L1 speaking population in Barcroft`s study would have likely had more cognates (Spanish L1 to English L2) than those of the present study (Japanese L1 to English L2). Finally, the time between exposure to the target vocabulary items differed greatly, as Barcroft looked at immediate post-test results following the lesson, possibly resulting in performance rather than learning. The present study investigated longer-term recall (three weeks and 12 weeks between vocabulary exposure).

In a smaller study that is more analogous to the present one, Iwata (2012) researched Japanese L1 high school students learning English L2 vocabulary using sentence generation. Iwata looked at one-week and four-week post-test results and found a significant improvement

in the generation group over the control group in both post-test time periods. The control group differed from that of the present study in that they focused on sentence translation from English to Japanese with the target vocabulary items embedded contextually. The present study`s findings fall between Barcroft`s (2009) and Iwata`s (2012), illustrating a need for further research using a larger population since this study`s knowledge scores showed a significant interaction suggesting a benefit of generation.

4.4 Limitations

This study has several limitations when considering the results. First, the small sample size may not represent the population at large. Additionally, although pairwise differences at each test point were not significant, Figures 4.2 and 4.3 show a trajectory of increased recall in interactions in the generation treatment that could possibly lead to a significant difference if the power from a larger sample size was used; however, this cannot be ascertained with the data collected in this study. Second, the participants and setting for the study were specific to B1 ELL first-year Japanese university students in Tokyo; hence, the findings may not be the same for L2 vocabulary learners in general. Third, the study only used a single measurement of vocabulary knowledge, meaning recall, which accounts for word explanation and sentence generation. Since the study did not account for other measurements (e.g., interlexical and intralexical factors), the single measurement should be considered a limitation. Finally, while

gender is a possible variable in L2 language learning in general (Iwaniec, 2019) and in vocabulary learning (Montero-SaizAja, 2021), this study did not run a covariate analysis for gender as was done in the Module 2 study discussed in Chapter 2. This was because the participant numbers in this study were reduced greatly due to Covid infections causing students to miss either the lesson, pre-test, or post-tests, leading to data not being used. The reduction in participant numbers, particularly among males, would likely not provide reliable findings. Data used in the analysis were only available from six males compared to 17 females, hence the lack of male participant scores would not allow for a reliable comparison. Gender has been shown in studies to have an effect, specifically where female learners have been found to be more successful (Mihaljević Djigunović, 1993). There is research directly related to the Japanese context where a social construct that females are naturally better language learners than males has been shown (Kobayashi, 2002). Kobayashi provides possible reasons for this social construct. The first is that the traditional marginalization of women in Japanese society may lead women to feel more empowered when using English since it is not part of traditional Japanese society. Another reason is that there are more professional choices and opportunities for women who are strong English speakers (e.g., working for a non-Japanese organization). These reasons may lead to motivation related to gender; therefore, the lack of gender as a covariate in this study is a limitation. Since generation has been shown that it may have a

benefit in L2 learning, the next chapter investigates generation in a different L2 context, reading comprehension.

5.0 Chapter 5: Study 2 Effects of Generation on Reading Comprehension

5.1 Introduction

Much of the literature on generation in the ELL classroom focuses on vocabulary retention and use, while less research has been conducted on generation and reading comprehension in the L2 context. This chapter investigates the role generation plays in reading among L2 learners. As discussed in Chapter 2, the role of generation in reading is largely built upon Wittrock`s generative process of comprehension model (Doctorow, Wittrock, & Marks, 1978) and the idea that reading comprehension can be improved (in L1 contexts) by active reading strategies (Wittrock, 1989; 1992). While the arguments are made and largely supported by Wittrock in the L1 context, the L2 reading context is not as clear. Therefore, this chapter describes a study that examines the effect that generation tasks have on short-term reading comprehension among L2 learners and investigates the use of both L1 and L2 as an output of generation. The present study was conducted as an action research study in the context of required Academic Communications classes at a university in Tokyo, Japan, in the fall of 2022 among Japanese L1 university ELLs. Within the curriculum of the course, there was a section on study skills requiring students to reflect on their L2 skills and learning habits. Reading in L2 was a core skill developed in the class. Therefore, students were required to practice and reflect on different study techniques to improve their reading ability in L2. The present action research study was conducted during the three weeks that focused on reading skills.

5.2 Reading Study Rationale

This study takes a different perspective on generation than the studies in Chapters 3 and 4 investigating the effects on reading comprehension in an L2 context. Additionally, it takes cognitive load theory into account by comparing not only generative and non-generative tasks but also specific generative tasks that may increase or reduce the amount of cognitive load during the activity. It compares two generation treatments with separate cognitive load levels to a baseline non-generation treatment. Some cognitive load proponents argue that desirable difficulties can become undesirable if they cause an overload in working memory capacity (e.g., Chen et al., 2015; 2016; 2018). Generation requires the learner to stretch their ability by completing an activity (e.g., reading task) and generating content (e.g., novel sentences, examples, or summaries). As the activities in this task require reading comprehension in L2, having the learner also produce content in L2 (i.e., higher cognitive load) may cause an over-taxation of cognitive resources, thereby reducing the effectiveness of the activity. Therefore, this study considers this by implementing a treatment where participants will read in L2 but produce in their L1 to reduce the burden on working memory (i.e., lower cognitive load).

It is important to note that using L1 in an L2 context has been argued to cause a splintering of attention from moving between languages on a given task, referred to as code-switching. Nawal (2017) found that using bilingual dictionaries in an L2 writing context may cause an increase in cognitive load from splitting attention between L1 and L2. On the other hand, Evans (2011), in a study on reading comprehension in L2, found that code-switching among L1 Korean university students did not cause a negative effect as students could better comprehend when using L1 alongside L2 to support their understanding of L2 texts. Bruen and Kelly (2017) also contend that using L1 in the L2 classroom can reduce cognitive load and

anxiety. They found, through interviews with L2 instructors teaching German and Japanese and among the students, that a balance of L2 with L1 for explaining complex terminology, grammar, and concepts can reduce cognitive load and better facilitate the L2 class.

The types of generation tasks used for comparison were descriptive explanations, summary recall, personalized descriptive connections, and opinion explanations. Descriptive explanations require the learner to explain in their own words a setting or situation (e.g., description of a path taken in a story in novel student-generated phrasing). Summary recall requires the learner to generate a summary of a reading or situation from memory in their own words. With personalized descriptive connections, the learner generates content where they are part of that content (e.g., if I were in that situation, I would …). Opinion explanations are used for the learner to generate and express their own opinion.

Two types of generation treatments were investigated: L1 generation, which is assumed to reduce cognitive load by asking participants to provide responses in their L1 (Japanese); and L2 generation, which is assumed to increase the cognitive load by requiring participants to complete generation tasks in the target L2 (English). Both treatments were compared to a baseline treatment that did not require the participants to use generation. While long-term learning is a key aspect of the efficacy of DDs, this study took a shorter-term view, focusing on immediate comprehension and understanding of reading from the perspective that generation during the reading process would lead to a deeper understanding of the reading task material. The specific questions this study looks to answer are: 1) Does generation enhance comprehension in an L2 reading context? 2) Does generation using L2 lead to different outcomes compared to generation using L1?

5.3 Methodology

5.3.1 Participants

The study`s participants consisted of 44 first-year university students in Japan in compulsory Academic Communication and Reading-focused courses meeting weekly in person. They were L1 Japanese speakers with an average age of 18, which is standard for first-year Japanese university students. All participants were ELLs estimated to be at the B2 CEFR level. As in Study 1 (Chapter 4), the university TEAP (Test of English for Academic Purposes) was used for class placement when the participants entered the university. The form of TEAP used only addressed listening and reading skills. Although part of the Academic Communication class curriculum, speaking and writing skills were not considered when the English placement level was administered.

The 44 participants in the study were in three groups (Group A = 17, Group B = 16, Group C = 11). Participants were allocated to groups by the class for which they were enrolled (e.g., Class A = Group A, Class B = Group B, Class C = Group C). Data from seven participants (Group A = 3, Group B = 2, Group C = 2) were removed from the analysis due to missing data from being absent on the day a lesson activity was conducted. If one lesson was missed, all data for that participant were omitted from the study. Data used in the analysis came from 37 participants who completed all tasks in the study. There were 18 females and 19 males. The participants majored in either engineering or business administration, with none majoring in a subject specific to English language content (e.g., English literature) which could provide more reading practice in English. Consequently, the groups` extent of English content exposure was generally limited to the course the study was conducted in, except for students who may

have studied English outside of the university curriculum during the semester when the study was conducted.

The participants did not receive compensation for taking part in the study, and the scores from the lessons were not calculated as part of the participants` grades, and this was explained at the beginning of the study and before each lesson task. While the scores were not reflected in the participants` grades, the activities were used as part of participation scores, and completing the lesson task was required. The fact that the lesson activities were part of the participation score was conveyed to the students, and if they engaged fully with the task, they would receive full points for the activity. At the beginning of the study, consent forms were provided to all participants in English specifying that students could opt out of having their scores included in the study (Appendix 1a). This was approved by the University of Birmingham Research and Ethics Committee.

5.3.2 Study Materials

This study included three reading lessons which were each followed by a quiz. The quizzes included ten reading comprehension questions made up of five True / False questions, and five Explanatory questions. The readings were from a publicly available ELL textbook (Nation, 2009). The vocabulary in the textbook was estimated to be at CEFR C1 level, which is one level higher than the estimated student CEFR level. Three readings from Nation (2009) were chosen, each approximately 300 words long. The pictures from the readings were not used, consequently only the text was present for the participants to ensure that answers were derived from reading without visual cues. Additionally, the reading text was slightly altered for the generation conditions. In both generation conditions, the reading was separated into three

parts, each including one or two prompts in text boxes for the participants to generate an answer. A separate empty box was provided for participants to write the generated answer to the prompt. Following the first prompt(s), the next section of text was provided, which also included prompts. The final prompt came at the end of the reading. For the baseline reading task, the entire reading was provided without prompts or boxes. The ten quiz questions were printed on the opposite side of the paper with enough space for participants to write their answers. Further details on the differences in treatments and how they were implemented are provided in the Procedure Section (5.5.4). The following reading units from Nation (2009) were used in developing the tasks and quiz questions: *The Fossil Hunters*, *The Mad Hatter*, and *The Tenacious Inventor*.

5.3.3 Rationale for Materials

This study was conducted in a classroom within a required university curriculum. Therefore, the time allotted during class for the study tasks was limited, and the types of materials used needed to be related to reading with the purpose of improving the students` reading skills. This was done by incorporating materials that were at the appropriate level to challenge the students, but not too difficult that the task lack significance. To ensure that participants were able to read the text, complete the treatments, and answer the comprehension questions, the reading task was limited to 300 words. This amount of text, along with the generation task, could be completed in approximately 20 minutes to 25 minutes, leaving enough time to complete the comprehension questions.

In the course, regular readings were done in 25-minute time blocks known as Pomodoro Sprints. The participants were accustomed to reading using this method for other class reading

tasks unconnected to the tasks in this study. The Pomodoro method is a focused activity where students concentrate on a single task for a set amount of time, often 25 minutes (Study Efficiently Using the Pomodoro Technique, 2021). Since the participants had used this method in previous class meetings during the semester, the instructor had a clear idea of how long it would take for students to complete the reading, treatment, and comprehension questions. Furthermore, a unit from the textbook where the study`s tasks originated was piloted in all three classes (Groups A, B, and C) two weeks before the start of the study. The piloted unit was not a reading used in the study but was at the same estimated level and length as the readings in the study. This confirmed that students could complete the reading and comprehension questions and that the content was challenging enough while not being overly difficult. The reading task's difficulty was calibrated to ensure participants actively engaged with the text for comprehension. Easy post-test questions or reading material would not accurately identify focused reading, potentially hindering the assessment of treatment efficacy. Conversely, overly challenging tasks could lead to disengagement or inability to answer post-test questions and limit data variability, reducing analysis power. Employing a CI CEFR level reading text one higher than the participants' estimated reading level was used to achieve this balance, which was confirmed by the self-reported focus during the pilot task. Additionally, the 300-word length of text was implemented so there would be enough reading content to implement embedded questions.

The comprehension questions consisted of two types: True / False and Explanatory questions. The five True / False questions provided a statement related to the reading. The participants would write true or false next to the statement. If the statement was marked false, a correction was required in a box provided on the page. For Explanatory questions, students

received an open-ended question (e.g., Why couldn`t Tim and Dean coexist peacefully?). Both True / False and Explanatory answers were to be in English. A two-pronged approach was employed to assess participant comprehension of the readings, utilizing both True / False questions for factual accuracy and Explanatory questions for a broader understanding of the specific text section. This dual method aimed to capture the nuances of comprehension and identify areas for improvement beyond a simple binary correct / incorrect classification. To keep the task at the 30-minute time limit, the post-test was designed to be concise with ten questions, divided between factual (True / False) and Explanatory questions. This was done to mitigate participant fatigue and ensure consistent effort across all questions. Employing a larger question set within the time constraint could have jeopardized the reliability and validity of the post-test answers.

5.3.4 Procedure

This study was conducted over a three-week period. Students were allowed to use a dictionary (paper-based or electronic) during the activity. The treatments included baseline, L1 generation, and L2 generation. The baseline treatment had the participants read the text twice without any extra actions. The second reading of the text was to account for the extra time the other treatments received in the reading. The L1 generation treatment had the students generate their unique responses to the reading section from the prompts provided in their first language (Japanese). The final treatment was L2 generation, where participants would generate their responses to the prompts in English. In both L1 and L2 generation conditions, the subsequent quiz questions were completed in English, as done in the baseline condition.

Participants received instructions before the reading and quiz not to start or look at the paper until instructed to do so. Additionally, they were told not to turn the paper over and look

at the quiz questions until they had completed the reading task and treatment. Once they completed the reading task and turned the paper over to the quiz side, they were not allowed to look back at the reading. The researcher, who was also the instructor, strictly enforced this. Furthermore, for the generation conditions (one treatment using L1 and another using L2), students were instructed not to return to the previous reading section once they completed the generation task for the section. The process had the participants read the first section and then answer the prompts completely for that section using the required treatment. They would then move on to the next reading section and prompt before doing the last of the three reading sections and prompt(s). Once finished with the readings, each participant would raise their hand, and the instructor would tell them to turn the paper over and complete the ten comprehension questions in English. This enabled the instructor to take a mental note of who was doing which part of the task and ensure that participants were not looking back at the reading. For the baseline condition, participants were instructed to read the entire reading twice before raising their hands to receive permission to start the quiz. The baseline treatment did not include prompts or boxes for writing on the reading side of the paper, just the reading itself. The quizzes for each reading were the same across all treatments. The activities took about 20 minutes for students to complete, but all students were given 30 minutes, enough time for all participants to finish the entire task and quiz.

All three groups, Groups A, B, and C, received the same reading topic and quiz during the same week. In week one, The Fossil Hunters topic was given to all groups, but the treatments differed for each group (Group A = L2 generation; Group B = baseline; Group C = L1 generation). This continued for each task so that all groups were exposed to each treatment

once. Table 5.1 shows the lesson topics and treatments for each week and group. The reading tasks for all three tasks with prompts and comprehension questions are provided in Appendix 4.

Table 5.1: Week and Topic for Each Group and Treatment

Week & Reading Topic	**Group A**	**Group B**	**Group C**
1 (The Fossil Hunters)	L2 Generation	L1 Generation	Baseline
2 (The Tenacious Inventor)	Baseline	L2 Generation	L1 Generation
3 (The Mad Hatter)	L1 Generation	Baseline	L2 Generation

5.3.5 Data Collection and Results Assessment

Post-reading comprehension tests were collected as soon as the participants finished to ensure that changes to the answers from looking back at the reading did not occur. Before scoring the post-tests, all three sets of treatment tests were combined and shuffled so that the treatment of the resulting answers would not be known to the marker. This was done to help eliminate any bias in scoring. This study compared the effects of three treatment conditions. Results included two types of data: binary for the True / False score data and non-binary with three levels (0, 1, 2) for the Explanatory score data. Scoring for the True / False questions was binary (correct / incorrect) since these questions had a clear answer. However, for Explanatory questions, partially correct answers were scored with half credit since these answers required a more comprehensive understanding of the reading section. Furthermore, some answers may have had the general idea correct in their answer but forgot specific details. This could be evidence of partial comprehension and, therefore, provides more nuance for data analysis. For example, if most students answered partially correctly, this may indicate a general comprehension but a lack of comprehension of specific details.

Two different linear models were constructed using Jamovi Project version 2.3, R version 4.1, and GAMLj (2019) to compare the effects of the two treatment conditions. The True / False scores were treated as correct or incorrect using a generalized linear mixed model with a binomial distribution. Explanatory scores used a linear mixed model. Data for the two groups were compiled in a single file for analysis. For both True / False and Explanatory scores, each question was included as one data point per item, with individual True / False and Explanatory scores at each testing point. True / False scores used "1" for correct or "0" for incorrect. For a False answer to be marked as correct, the incorrect statement in the question needed to be corrected. For example, if the True / False question stated, "Mt. Fuji is the highest peak in the world.", to be considered correct, the answer would need to include the following information, "False, Mt. Everest is the highest peak in the world." If False was chosen correctly and the correct answer was provided, but the entire sentence was not re-written, the question was also scored as correct, such as "False, Mt. Everest."

For Explanatory question scoring, three scoring options were used (0 = completely incorrect or no answer given, 1 = partially correct, 2 = correct). Grammar and spelling mistakes did not count against the score being marked correct. If the information in the answer was clear enough to understand and correct or partially correct, it was scored as such. The True / False questions and Explanatory questions each included five questions for each of the three reading tasks (5 questions x 3 tasks = 15 data points per participant for T/F and 15 data points per participant for Explanatory). There were 37 participants in total from all groups that completed the study, which gave an overall total of 555 data points for each question type.

5.3.6 Results in Aggregate for True / False Questions

The overall results of the True / False question scores are included in Table 5.2.

Table 5.2: Mean (standard deviation) and 95% Confidence Intervals (CIs) for Scores (/1) on True/False Questions

Condition	True / False Scores (/1)	95% CIs
Baseline	0.68 (.47)	0.61, 0.74
L1 Generation	0.65 (.48)	0.58, 0.72
L2 Generation	0.59 (.49)	0.52, 0.66

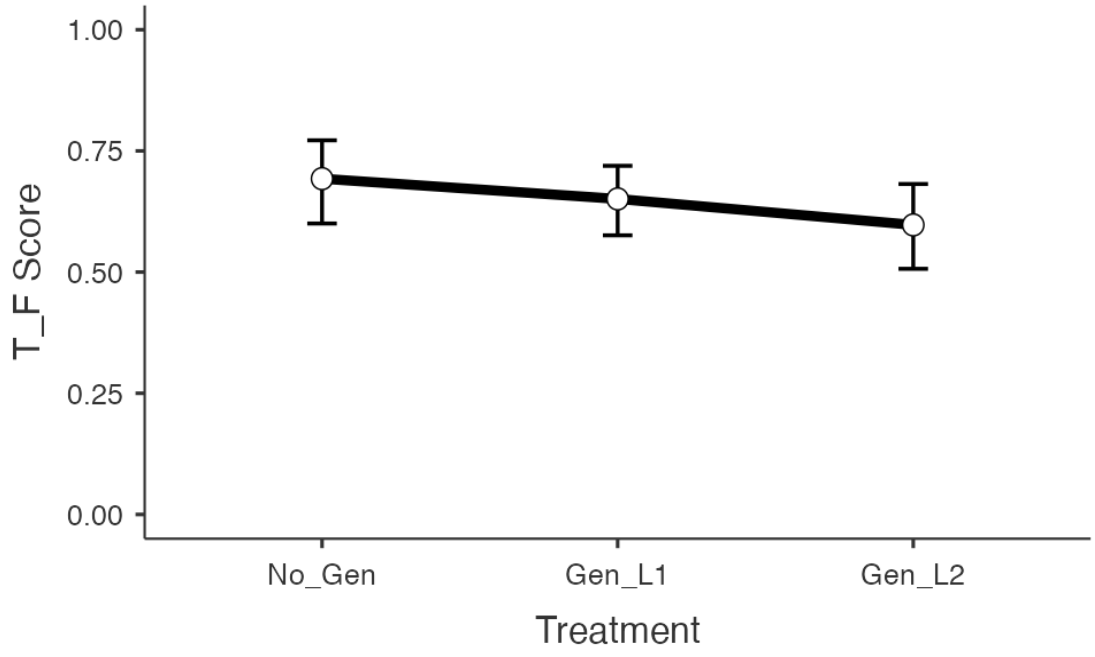
An initial model included fixed effects of condition, treatment coded with baseline as the reference level. Random intercepts for subject and by-subject random slopes for the effect of treatment condition were included.[2] The model investigated L1 and L2 generation treatments. The True/False scores did not show a significant effect for generation in L1 (β = -0.190, z = -0.735, p = .462) nor in generation L2 (β = -0.418, z = -1.363, p = .173), see Table 5.3. Post Hoc test pairwise analysis, with Bonferroni correction for multiple comparisons, confirmed that L1 generation (z = 0.735, p = 1.00) and L2 generation (z = 1.363, p = 0.519) were not significantly different to the baseline condition. In addition, there was no significant difference between L1 generation and L2 generation conditions (z = 0.873, p = 1.00). Figure 5.1 compares the treatments` True / False scores.

[2] As there were no "items" in the same way as in the vocabulary study, these were not included in the random effects structure for this analysis.

Table 5.3: Model Output for the Effect of Treatment (Reference Level = Baseline) on True / False questions

					95% Exp(B) Confidence Interval			
Names	**Effect**	**Estimate**	**SE**	**exp(B)**	**Lower**	**Upper**	**z**	**p**
(Intercept)	(Intercept)	0.610	0.0964	1.840	1.524	2.22	6.329	< .001
Treatment1	Gen_L1 - No_Gen	-0.190	0.2579	0.827	0.499	1.37	-0.735	0.462
Treatment2	Gen_L2 - No_Gen	-0.418	0.3069	0.658	0.361	1.20	-1.363	0.173

Figure 5.1: Side by Side Comparison of Treatments, including 95% Confidence Intervals (CI)

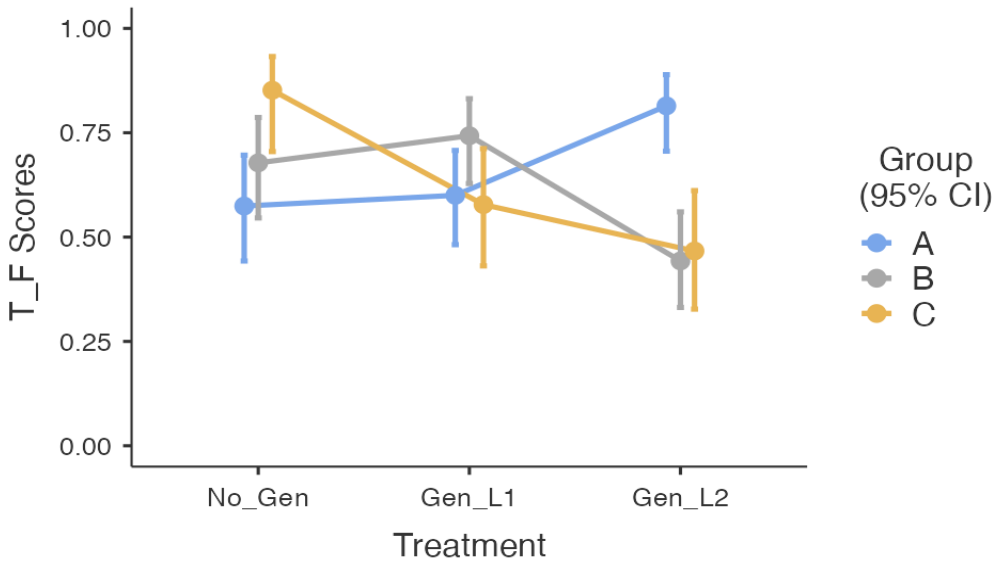


5.3.7 Covariates: Gender, Group & Lesson for True / False Questions

Gender, Group, and Lesson were then added to the base model as covariates (see Appendices 4a, 4b, and 4c for full output of these models). Gender was not significant either as a fixed effect or as an interaction with condition.

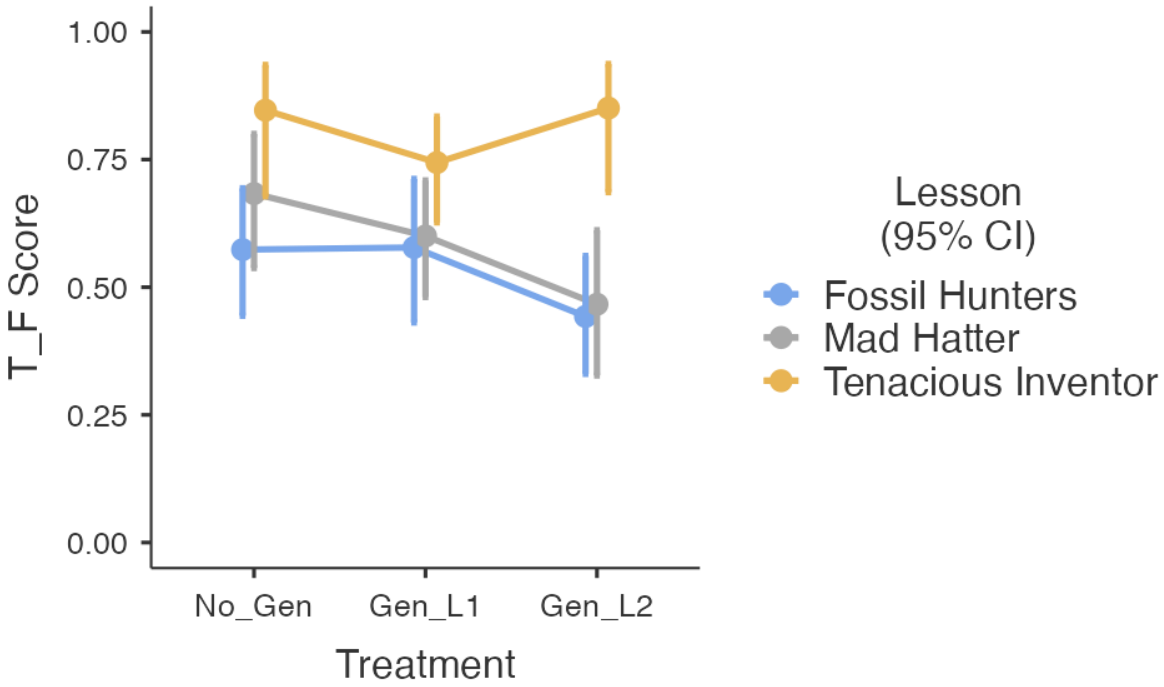
For Group, the overall differences between A and B (z = -0.922, p = .357) and A and C (z = -0.33, p = .742) were not significant. There was a significant interaction between Groups (A-B) and condition (L2 generation): z = -3.86, p = < .001; Groups (A-C) and condition (L1 generation): z = -2.38, p = .017; Groups (A-C) and condition (L2 generation): z = -4.51, p = < .001. As shown in Figure 5.2, while Group C (and to some extent Group B) showed a cost for L2 generation (compared to the baseline), Group A did not show the same cost.

Figure 5.2: Interactions of Groups and Treatments for True / False Questions, including 95% Confidence Intervals (CI)



For Lesson, there was a significant effect for the lessons The Tenacious Inventor and Fossil Hunters ($z = 5.08$, $p = < .001$). As shown in Figure 5.3, The Tenacious Inventor lesson was overall completed more successfully, and did not follow the same pattern as the others for the L2 generation treatment.

Figure 5.3: Comparison of Lessons and Treatment Scores of True / False Questions, Including 95% Confidence Intervals (CI)



Generally, L2 generation came at a cost, but the additional analyses demonstrate that this may depend on factors such as lesson difficulty and group ability.

5.3.8 Results in Aggregate for Explanatory Questions

The overall results of the explanatory questions are shown in Table 5.4.

Table 5.4: Mean (standard deviation) and 95% Confidence Intervals (CIs) for Scores (/4) on Explanatory Questions

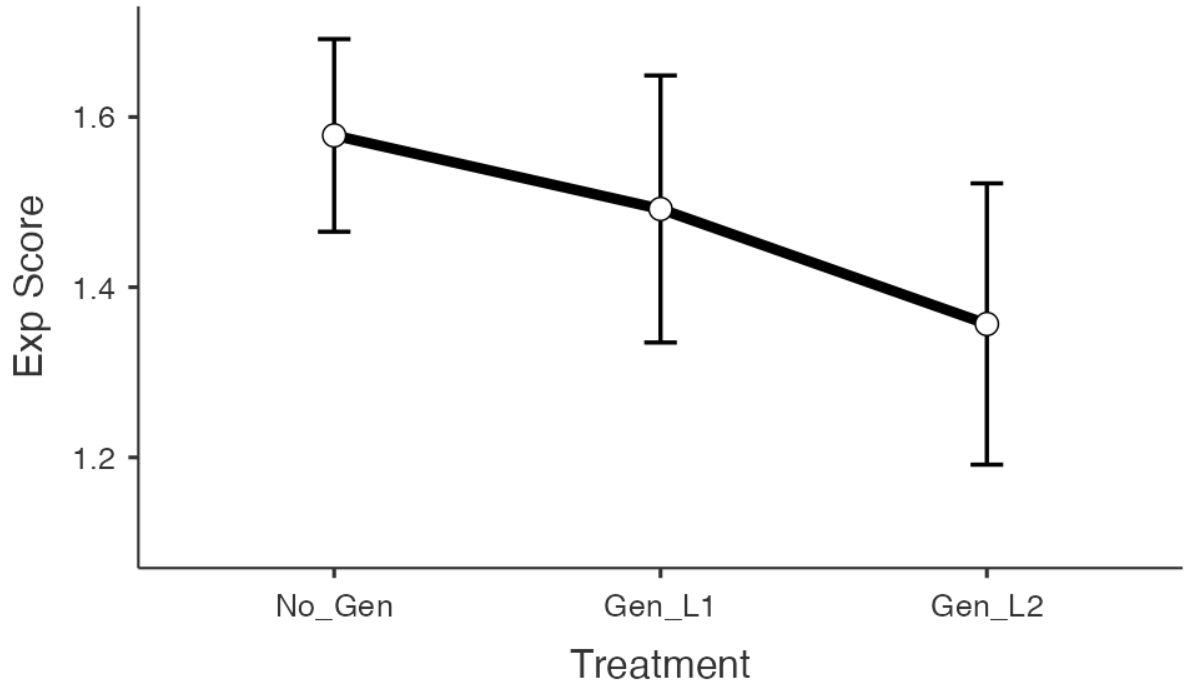
Condition	Explanatory Scores	95% CIs
Baseline	1.58 (.66)	1.48, 1.67
L1 Generation	1.49 (.79)	1.38, 1.61
L2 Generation	1.36 (.86)	1.23, 1.48

An initial model included fixed effects of condition, treatment coded with baseline as the reference level. As in analysis for True / False questions, random intercepts for subject and by-subject random slopes for the effect of treatment condition were included. As with the True / False scores, the Explanatory scores did not show a significant effect for generation in L1 (β = -0.087, t = -0.976, p = .334), but there was a significant difference for generation in L2 (β = -0.222, t = -2.622, p = .011), see Table 5.5. Post Hoc test pairwise analysis, with Bonferroni correction for multiple comparisons, confirmed that L1 generation was not significantly different to the baseline condition (t = 0.976, p = 1.00), but L2 generation was significantly lower than the baseline (t = -2.62, p = .038). In addition, there was no significant difference between L1 generation and L2 generation (t = 1.484, p = .439). Figure 5.4 compares the treatments` Explanatory scores. L2 generation led to significantly lower scores overall than the baseline condition.

Table 5.5: Model Output for the Effect of Treatment (Reference Level = Baseline) on Explanatory Scores, z and p-values with Baseline and Confidence Intervals

				95% Confidence Interval				
Names	**Effect**	**Estimate**	**SE**	**Lower**	**Upper**	**df**	**t**	**p**
(Intercept)	(Intercept)	1.4757	0.0520	1.374	1.5776	36.4	28.381	< .001
Treatment1	Gen_L1 - No_Gen	-0.0865	0.0886	-0.260	0.0872	44.6	-0.976	0.334
Treatment2	Gen_L2 - No_Gen	-0.2216	0.0845	-0.387	-0.0560	65.0	-2.622	0.011

Figure 5.4: Side by Side Comparison of Treatments for Explanatory Scores, including Confidence Intervals (CI)

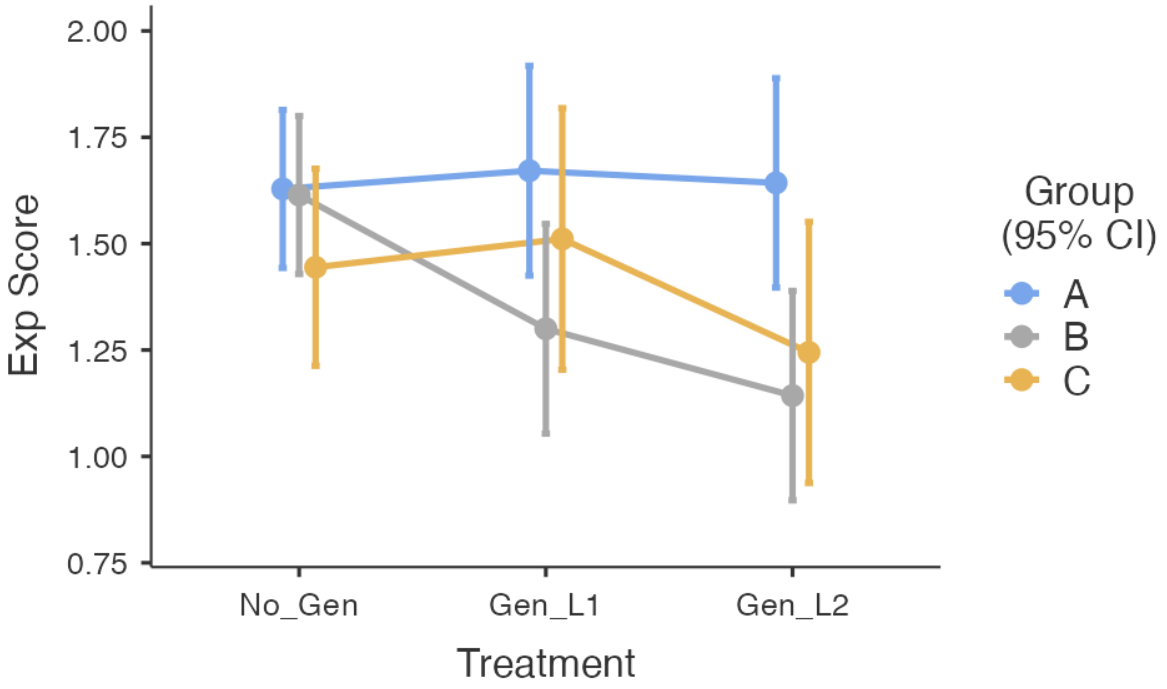


5.3.9 Covariates: Group, Gender & Lesson for Explanatory Questions

Gender, Group, and Lesson were then added to the base model as covariates (see Appendices 4d, 4e, and 4f for the full output of these models). Gender was not significant either as a fixed effect or as an interaction.

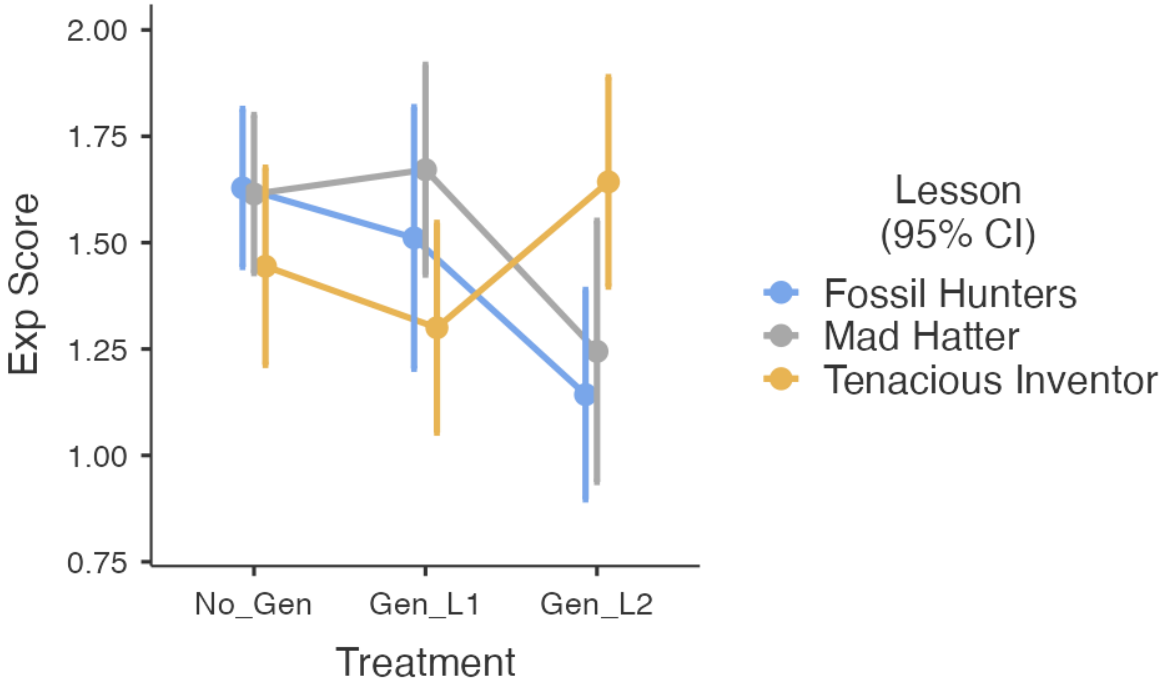
For Group, the overall difference between A and B was significant (t = -2.67, p = .011), but the difference between A and C was not (t = -1.98, p = .056). There was also a significant interaction between group (A-B) and condition (L2 Generation): t = -2.64, p = .010. As shown in Figure 5.5, while Group B (and to some extent Group C) showed a cost for L2 generation compared to the baseline condition, Group A showed minimal differences across the three conditions.

Figure 5.5: Covariate Group Interactions with Treatments for Explanatory Scores, Including Confidence Intervals (CI)



For Lesson, there was a significant interaction for the lesson Tenacious Inventors and L2 generation (t = 2.83, p <.001). As shown in Figure 5.6, The Tenacious Inventor Lesson did not follow the same pattern as the others, especially for the L2 generation condition.

Figure 5.6: Covariate Lesson Interactions with Treatments, Including Confidence Intervals (CI)



Overall, as in the T/F question analysis, L2 generation in general came at a cost, but the additional analyses demonstrate that this may depend on both group ability and lesson difficulty.

5.4 Discussion

Data obtained in previous generation studies on reading comprehension and recall in L1 contexts (e.g., Begg et al., 1989; DeWinstanley & Bjork 2004; Wittrock & Alesandrini, 1990)

have shown mixed results in using generation in reading activities. The present study examined the effect of generation on L2 reading comprehension. This was done by examining the effects of two generation treatments compared to a baseline condition without generation. The two generation treatments using L1 and L2 were also compared with each other to provide insight into possible variable effects of more / less cognitive load. Other studies that have looked at the use of L1 in an L2 context (e.g., Evans, 2011) allowed for L1 use through dictionaries or asking questions for clarification but did not require L1 as a part of a generation task as this study investigated. In this section, the results are discussed with reference to the two questions that the study attempted to answer.

5.4.1 Question 1: Does generation enhance comprehension in an L2 reading context?

The present study failed to show a benefit or cost of L1 on reading comprehension in the L2 context compared to the baseline; however, L2 generation showed a clear cost compared to the baseline. In line with the results from Abel and Hänze (2019), generation in the present study failed to show a benefit in reading comprehension for L2 contexts when tested immediately following the task. The lack of DD efficacy in short-term performance explained by Soderstrom and Bjork (2015), where DDs hinder immediate performance, was seen to some extent in this study. As cognitive overload pulls on the required attention of a particular task, the costs seen from the generation task in L2 could result from attention to the reading text being disrupted. A possible explanation for this is cognitive overload, which occurs when the processing capacity of cognition is exceeded. One important consideration in the design is that for the baseline condition, all participants who were not in the generation treatment conditions, read the text twice. The rereading of the text was implemented to balance the conditions in

terms of time on task, and the generation condition failed to show a benefit over the baseline condition.

Reading the text twice was used for two reasons. First, the time needed to complete the generation tasks exceeded reading the task once through. Unfortunately, a follow-up test after time had passed (e.g., three weeks) was not conducted to verify if this was the case. This was due to the timing of university holidays making it so each group would not have equal time between proposed follow-up post-tests. The second reason for rereading as the baseline condition is that this is a common study practice among university students (Kirschner & Hendrick, 2020). However, the literature on DDs and in the learning sciences, in general, does not support rereading for learning (Brown et al., 2014; Dunlosky, Rawson, Marsh, Nathan & Willingham, 2013; Soderstrom & Bjork, 2015). Nevertheless, there was a lack of any significant benefit from generation in either language (L1 or L2) in this study, and in one case, generation using L2 showed a significant deficit compared to the baseline. This may be attributed to an increase in cognitive load from generation. This can be surmised because generation in L2 included the most complexity and, therefore, the highest burden on cognitive load. Conversely, rereading was the least complex treatment and, therefore, is expected to have incurred the least burden on cognitive load. L1 generation did not show a benefit over the control, but it also did not show the same cost seen in L2 generation. This may be because L1 generation may have increased the burden on cognitive load, but not enough to cause a deficit. Conversely, difficulty added was not sufficient to lead to any increase in comprehension, at least as measured here in an immediate test.

5.4.2 Questions 2: Does generation using L2 lead to different outcomes compared to generation using L1?

The effect that generation condition caused on learners` cognitive load could not be definitively established in this study. The inclusion of an L1 generation condition was designed to assess differences in cognitive load since it was assumed that this would present an easier task than generation in the L2. If L1 performed significantly better than L2, this would have provided support for the idea that L2 may have overtaxed learners by increasing cognitive load to unhelpful levels. However, in this study, L1 generation did not perform significantly better than L2. Nevertheless, Figures 5.1 and 5.3 show a clear downward trend as the complexity and, therefore, likely cognitive load increased. Additionally, in the analysis for Explanatory questions, there was a significant disadvantage for L2, but not L1, compared to the baseline condition. A possible reason that L1 did not significantly improve scores is that the tasks where L1 generation treatment was used led to split attention. Attention may be split when an L2 learner uses L1 in the same activity, causing a code switch between languages. What is not clear in the results of this study is whether attention splitting produces enough of a cost to hinder any advantage while not enough to cause a disadvantage to the baseline treatment. Additionally, interspersing generation tasks in both generation treatments in the reading task may have caused split attention as it requires the learner to go from the task of reading to generating, then back to reading. Although not found in the present study, using embedded questions in an L1 context has been shown to assist in comprehension (Walczyk & Hall, 1989), and in L2 reading speed (Hung, 2009). Either, or a combination of both split attention factors, may have had a negative effect on the generation treatments, canceling out any benefit.

However, the results from the data gathered cannot support or refute this possibility definitively.

5.4.3 To what extent did covariates affect the findings?

This study looked at gender, group, and lesson as covariates. Gender did not have any effect in either set of questions, nor did it have an interaction with treatments. Gender was included as a covariate as research among language learners shows findings of a possible effect (Iwaniec, 2019; Montero-SaizAja, 2021). Group showed some differences in interactions between groups and L2 generation. This interaction is shown in Figures 5.2 and 5.5. Lesson showed that The Tenacious Inventor lesson did not follow the same pattern as the other lessons for the L2 generation treatment. This was found in both sets of questions. Possible reasons for this difference could be the result of the lesson The Tenacious Inventor being less challenging or more familiar to the students. Therefore, if generation overloaded the participants, the role on cognitive load seems to depend on lesson difficulty and group ability, and clearly the overall L2 cost is not always the case. Thus, factors of lesson difficulty and group ability need to be considered.

5.5 Limitations

This study has several limitations that should be considered when interpreting the results. First, the sample size was relatively small and may not represent the larger population. Second, the participants and setting for the study were specific to B2 ELL first-year Japanese university students in Tokyo, and the majors the students were part of were limited to engineering and business administration. Finally, the study was conducted in a classroom as

part of a set university curriculum, which may include outside factors (e.g., social clubs and other class exams) that may have affected the students` attention on the specific days present study`s tasks took place. Despite these limitations, the study provides valuable insights into the effects of interspersing generation in L1 and L2 on reading comprehension. The small sample size was due to class size constraints comprising the groups. While the results may not be generalizable to settings outside of Japan, they provide a starting point for future research. Future research could address the limitation by utilizing a larger sample and collecting data from participants in an intensive language program where participants might better represent a wide range of L1s.

In addition to the researched questions, other questions remain; for example, would increasing the time between the reading and the comprehension questions lead to different results? In other words, was the reading done in the baseline task too fresh in the participants' minds and therefore made answering the questions relatively easy? This was not addressed formally in the study since the concern was whether generation use would benefit comprehension. Another question remains: is there an effect of adding extra difficulty through generation on different English language abilities (e.g., A2 or C1)? Therefore, conducting studies comparing different ages and L2 levels would also be advantageous. Additionally, time on task, in both reading of the text and time taken to answer questions, could be further explored. Finally, this study incorporated various generation activities interspersed in the text: descriptive explanations, summary recall, personalized descriptive connections, and opinion explanations. These can be researched independently to explore aspects of generation independent of each activity.

5.6 Conclusion

To my knowledge, the study reported here is the first to investigate cognitive load aspects requiring the use of L1 generation in an L2 reading context. Furthermore, this study focused on generation in both L1 and L2, differentiating it from previous research. The findings suggest a lack of benefit of generation on reading comprehension, but the data do not give clear reasons for this. Nevertheless, there are similarities to other studies that identified cognitive load issues playing a part in an L2 context, and in L1 context. The results follow similar negative effects on vocabulary learning when using generation (e.g., Bancroft, 2006) and did not show benefit in comprehension from using embedded questions in the text. In Chapter 4, generation showed some interactions in vocabulary recall; however, similar benefits were not seen in this present study in reading comprehension. In the next chapter, Chapter 6 will look at generation in the L2 writing context.

6.0 Chapter 6: Study 3 GENERATION CHECKLISTS & WRITING

6.1 Introduction to Writing Checklist Study

This action research study investigated the use of hybrid post-writing and reflection checklists as a generation task in the L2 writing context. It examined whether implementing self-feedback generation with reflection reduces mistakes and improves writing content. The present study presented a set of writing reflection and post-writing proofreading questions embedded in a writing checklist. Findings based on the scores comparing students who undertook these activities with or without generation (e.g., with or without being asked to explicitly produce content to support the checklist) will help provide insight into their efficacy in an L2 writing context. The questions the study set out to answer were: 1) Does the use of generation in post-writing reflective checklists enhance academic writing in the L2 context? 2) Are there differences in efficacy for generation tasks focused on content versus editing (formatting and GSPC)? 3) Is there a carryover effect of prior generation use on subsequent writing tasks?

6.1.1 Checklists

Student reflection on their writing can serve multiple purposes, and reflection can and usually does, take place to different extents throughout the writing process. This includes times when the writer wants to focus on a specific writing task (e.g., editing) but finds that they are still reflecting on other aspects of their writing despite the implementation of scaffolding and process writing. This difficulty in the writing and revising stages might lead to mistakes (e.g., punctuation and grammar). The use of checklists may benefit the writer by providing some

support to help alleviate some of the demands on working memory during a task (Gawande, 2011; Schiano, 2021), much as a checklist used for a later draft in proofreading may be able to serve as a safety net during the writing process. If writers know that they will have the opportunity to fix any final mistakes, they might be more comfortable focusing on the writing stage (e.g., the Carpenter stage in the Flowers Paradigm) and not reflecting on editing. While there is evidence for using checklists as a safety measure not to miss a step in a procedure (see Gawande, 2011), there is not much support for or against their use in L2 writing.

Reflection in writing implements multiple DDs, such as retrieval, elaboration, and generation (Brown, Roediger & McDaniel, 2014), to help students monitor their writing and move from novice to skilled writer. When a writer can reflect on their own writing, it shows autonomy, a requirement for skilled writers to progress outside of the classroom. One example of how self-reflection is often used in L2 writing is through reflective journaling. This is a low-stakes way of looking at one`s writing that can solidify skills or content and assist in self-regulated learning (Nückles et al., 2020). Because of the low-stakes nature, often not graded or read by anyone but the writer, there can be much more focus on content during the reflective writing task. This is especially important for novice writers who struggle more with the mechanics of writing, which may take more attention away from content in higher-stakes writing.

In other tasks where the monitor is more involved, and the stakes are higher, the use of DDs, such as metacognition through self-feedback, may cause cognitive load issues (Cadaret & Yates, 2021; Chen et al., 2015; 2016; 2018). Since reflection in writing integrates multiple DDs, the chance of increased stress on cognitive load could be exacerbated. Nevertheless, according to Hyland (2019) students need to do cognitively challenging tasks

to develop writing. While reflection elicits DDs, the checklist itself may reduce the amount of cognitive load if it is used in later drafts or as part of a writing process the writer trusts. Trusting the process would be necessary for the writer to keep attention on the current writing task. Without trust in the process, the writer may be less likely to ignore thoughts related to different tasks in the process, thus leading to an increase in cognitive load from the split-attention effect. Jagaiah, Howard, and Olinghouse (2019) describe writing checklists as a tool that can be used as a procedural facilitator, supporting students through each stage of the writing process, in addition to reminding the learner what they need to edit after writing. Additionally, for L2 learners to become independent writers in the target L2, they must take on the responsibility of editing mistakes (Hinkel, 2020).

Post-writing checklists assist the student and instructor in identifying gaps in student understanding, a positive aspect of DDs (Soderstrom & Bjork, 2015). These checklists focus on final proofreading and editing, not revision. Revising involves expanding or changing major aspects of the draft, while editing is the final stage of the process. Here grammar, spelling, and other similar aspects are corrected (Jagaiah, Howard, & Olinghouse, 2019). A key benefit of DDs, such as metacognitive feedback, is illustrated in the mistakes not corrected after this post-writing checklist is completed. Uncorrected mistakes provide insight to the instructor about a possible gap in learning and help to prevent false fluency.

In addition to checklists used for writing procedures and proofreading, checklists can be used as a reflection tool incorporating self-feedback, as discussed earlier in this section. Nielsen (2014) explains that checklists can help students focus on a specific aspect of writing depending on their level. She further argues that they are effective reflection tools that encourage the student writer to reflect on their writing and analyze their writing from a

figurative distance with a clearer perspective. Self-reflection post-writing checklists require the writer to think about their writing after completion. While questions on grammar, spelling, formatting, and other specific points can be included, the purpose is to get the student to think about what they do and do not understand by looking deeper at their writing and what they learned from the task. Camp (1998) argues that reflective questions should be used on specific pieces of writing and writing portfolios to help develop an awareness of the writing process being used. This awareness requires the writer to retrieve what they did in the writing process and to elaborate, which is another form of generation.

6.2 General Rationale

6.2.1 Writing Checklist Lessons and Treatments

This study was a classroom action research project closely aligned with the class syllabus and university curriculum. Therefore, the number of tasks and the types of writing were predetermined in the course curriculum. The two writing checklists, control (non-generation) and generation, covered the same points but differed in the amount of extra reflection required. In the control treatment, a list of points to check by the participants was included, whereas in the generation treatment, a written reflection on each point and a general reflection paragraph on the essay as a whole were used. The points included Formatting, GSPC (Grammar, Spelling, Punctuation, and Capitalization), and Content. These three aspects of writing were separated according to perceived cognitive difficulty.

Formatting included proper margin width, centering of the title, using the correct font size and style as required by the university writing standards, 2.0 line spacing, and indenting of new paragraphs. This closely followed APA (American Psychological Association) Seven

guidelines. As a step-by-step guide and examples were provided to all students on how to format papers correctly, this was expected to be the least cognitively taxing of the three areas covered by the checklists since students could see the guide and example as they wrote their assignments.

GSPC covered standard writing expectations in grammar, spelling, punctuation, and capitalization, which students were expected to be able to understand and use correctly at the class level in which they were placed (B1 CEFR). While spelling and capitalization are more clearly defined and were expected to be mistake-free since students used spell check in Microsoft Word, grammar and punctuation were considered slightly differently. In this study on writing, mistakes are defined as something done incorrectly when the writer knows (or should know) how to do it correctly. Since grammar correction is not as clearly defined as spelling and capitalization mistakes, these two aspects were treated differently. Errors are defined in this thesis as something done incorrectly when the writer does not know or is not expected to know how to do it correctly (Ellis, 1997). For example, if a student uses the future tense in a sentence describing something that happened in the past instead of the past tense, this would be considered a mistake by the writer since they likely know or are expected to know this grammar skill at the B1 level. Conversely, if the writer tried to use the past perfect tense but chose the wrong verb form, this was considered an error because this was not yet learned or is considered more advanced and beyond what is expected of the student at the current class level. Mistakes were considered wrong and received negative scores, while errors were not considered in the study's scoring. Punctuation was similarly scored as grammar; hence simple full stop and question mark problems were considered mistakes. In contrast, more advanced punctuation, e.g., in-text citations done incorrectly, was considered an error for early tasks

(Essays 1 and 2). As in-text citations and other new punctuation or writing aspects were covered during the course, subsequent writing task scoring changed as what was considered an error before being taught in the class became a mistake after being taught. GSPC parts of the essay were considered the second most cognitively taxing of the three areas scored.

The third area covered and scored by the checklists was that of content, which was considered the most cognitively taxing of the three. Content spanned the length of the writing task (e.g., meeting minimum assigned word count), use of proper essay structure and logic (e.g., thesis statements, topic sentences, concluding sentences, and supporting sentences), following the assigned topic and essay type of the task, and the content of the sentences (e.g., word choice). Content required the most reflection with the least amount of clear modeling of the three skills tested in this study. While examples were provided, the specific content of each student`s essay differed from that of the model provided in the student textbook. Additionally, in Essays 3 and 4, students used source material unique to their essays. Finally, Content did not have the assistance of Microsoft Word spell-checking or grammar checking. Likewise, the specific steps provided in Formatting from the model guide were not provided for Content, requiring more depth of thought.

Many points in the post-writing checklist were purposefully broad, as the checklist`s goal was to ensure that students proofread the essay holistically and reflect on their writing instead of just checking boxes as done. However, this is what some students may have done in the control checklist, as there was no way to confirm reflection on each point beyond what the scoring revealed. The complete checklists used in both control and generation treatments are provided in Appendices (6a = Generation, 6b = Non-Generation).

6.3 Methodology

6.3.1 Participants

The participants in this study consisted of 39 first-year Japanese (Japanese L1) university students in an ELL academic reading and writing class, the second of two compulsory reading and writing classes for first-year students at the university. The students were, on average, 18 years old and majored in economics, finance, or business. Prior to starting classes in the spring semester, the participants took a university-developed English placement test which placed them at the CEFR B1 level. The placement included reading and listening assessment, but it did not include writing as part of the placement test.

The two classes, henceforth referred to as Group A and Group B, were held on Saturday mornings during the fall semester in a face-to-face on-campus setting. Group A was the control group, and Group B was the generation group. Group A had 19 participants, and Group B had 20 participants at the beginning of the class. The classes met 14 times, with each session lasting 100 minutes. Data from eight participants were not included in the final analysis due to participants not completing tasks. At the end of the study, data from 31 participants were used in the analysis (Group A = 14, Group B = 17). There were 21 females and 10 males (Group A = 6 males, 8 females; Group B = 4 males, 13 females). Gender was not considered in this study as a covariate, as there was a lack of male data. The participants were well matched as far as age, university majors, English placement, and day and times of classes studies were conducted (Saturday mornings). One aspect the groups were not well matched was Gender.

This study did not provide compensation to participants. However, the checklist treatments were used as part of the final essay grade in that if they were completed on time and included with the final essay on Blackboard, the learner management system used in the class,

they received points for the checklist. The differences in treatments did not alter the grading of the checklists. Participants were provided with a consent form detailing that they could opt out during the study from the data being used in the analysis by contacting the instructor (Appendix 1a). This project was approved by the University of Birmingham Research and Ethics Committee.

6.3.2 Study Materials

The study used the following materials in the course to develop class writing submissions: A publicly available textbook on English reading and writing skills (Daise & Norloff, 2019) and Blackboard learning management system. The classes included weekly textbook reading and comprehension tasks in the lead-up to each unit writing task. These writing tasks were used in the treatment assessments. The four writing tasks in the study were Essay 1 (Summary writing), Essay 2 (Cause and effect), Essay 3 (Opinion with references), and Essay 4 (Persuasion with references). An electronic version of the checklist for both treatments was provided on Blackboard as a Microsoft Word document (Appendices 6a, 6b).

6.3.3 Procedure

This study was conducted over 14 weeks, the entire length of the fall 2022 semester. The first two weeks of the study consisted of textbook reading activities, starting with Unit 5 (of 10) textbook units. The courses continued from where the spring semester ended with the completion of Unit 4 in the reading textbook (Daise & Norloff, 2019). As different instructors taught the two classes (Groups A and B) in the spring semester from the instructor in the fall semester, the exact amount of content covered in the textbook could not be verified. Therefore,

both groups started at the same place, Unit 5, of the textbook. In the second week of the course, the first essay assignment was introduced. Students were assigned to complete a draft of the Summary writing task for homework. This followed a model example of what was expected as provided by the textbook. Furthermore, specific guidelines on the formatting of the essay were provided along with examples for student reference. The assigned typed draft was brought to class for peer review in Week 3. Peer review required students to check each paper in their peer-review group (approximately four to five papers). The instructor signed each draft to ensure completion of the assignment and monitored the groups to make sure all group members were participating actively in checking each other`s essays. Students were allowed to speak in English or Japanese during this activity.

Following the first peer review, the checklist was introduced to the class through PowerPoint slides, and instructions on completing the checklist were given according to the specific treatment of the class (Group). Group A served as the control group and was instructed to mark each checkpoint one by one as they verified that they had correctly done this in the essay. Checkpoints were only to be checked after any corrections were made or it was confirmed that corrections were not needed. Group B served as the generation condition and received the same initial instructions as Group A. However, Group B was instructed to generate an explanation of why they were confident they did it correctly or what mistakes they found when revisiting the checklist point in their essay. They were also to include any questions they were still unsure about. Students were expected to write at least two or three sentences reflecting on the point but were encouraged to write more. A model of the type of reflection expected for each point was provided at the beginning of the checklist and explained during the

in-class PowerPoint directions on how to complete the checklist task. A model reflection generated answer is provided in Figure 6.1.

Figure 6.1: Model Example of How to Write a Generated Answer for Each Checklist Point, Provided to Generation Condition Group

Example:

(√) I checked to make sure I used the correct tense in my writing.

I reread my essay and found some sentences had verbs in the present tense, but I was writing about the past tense. I am confident it is correct because I reread the essay and checked each verb tense. Most of my mistakes around verb tense happened with "to be" verbs like "are" and "were".

At the end of the generation treatment checklist, participants were required to generate a paragraph reflecting on their writing after completing the checklist. The exact instructions were provided on the checklist. Additionally, during the PowerPoint checklist explanation for the generation treatment, students were told to write a complete paragraph of at least seven sentences but were encouraged to write much more without any length limit. Figure 6.2 provides the final reflective paragraph instructions for participants in the generation condition.

Figure 6.2: Instructions for Final Reflective Paragraph for Generation Condition

Instructions for completing the final reflective paragraph:

Write a short paragraph explaining what points you did correctly and incorrectly in your writing. Specify which of the above points gave you trouble and how you went about fixing them. At the bottom of the paragraph, type how long it took you to complete the checklist and write the checklist paragraph (for example, 30 minutes).

In Week 4, students submitted the following to the instructor: the first peer-reviewed Summary draft, the completed checklist for the final draft, and the final Summary essay draft. Moreover, this provided a clean copy of the essay for scoring for the purposes of the study. The subsequent assignments followed a similar schedule. Table 6.1 provides a week-by-week description of in-class activities, writing assignments, and submissions.

Table 6.1: Weekly Schedule of Writing Tasks and Submissions

Week	In-class Writing Activity	Homework Writing Task	Submission for Scoring
1			
2		Summary Essay (draft 1)	
3	Summary Peer-review	Summary Essay (Checklist, final draft)	
4			Summary Essay Checklist, and final draft
5		Cause & Effect Essay (draft 1)	
6	Cause & Effect Peer-review	Cause & Effect Essay (Checklist, final draft)	
7			Cause & Effect Essay, Checklist, and final draft
8		Opinion Essay (draft 1)	
9	Opinion Peer-review	Opinion Essay (Checklist, final draft)	
10			Opinion Essay, Checklist, and final draft
11		Persuasion Essay (draft 1)	
12	Persuasion Peer-review	Persuasion Essay (final draft)	
13	Final exam on reading & vocabulary from textbook given this week. An extra week was given to submit final Persuasion Essay.		
14			Persuasion Essay, final draft (No checklists for either group were used for the final essay)

6.3.4 Data Collection and Assessment

Data were collected from the four completed final writing tasks. Scanned photocopies of each final submission and the corresponding completed checklist were made before providing feedback and a grade on the task to the students. The scores used in the study were not the same as the grades given to tasks the students received. The student essay grade was scored out of 30 points, which included assessment for completing assignments (peer review draft, checklist, final draft) on time, participating in peer review, properly completing the checklist, and the writing content of the essay. Student writing task grades were more correlated to completing all submissions correctly and on time, which included properly completing the checklist. However, the scoring used for the study was specific to Formatting, GSPC, and Content. Each essay received a score of one, two, or three points for Formatting, GSPC, and Content. Scoring reflected the following: one = "did not meet assignment expectations", two = "met assignment expectations", three = "exceeded assignment expectations".[3] Additionally, under both treatments, students were instructed to mark the time it took to complete the reflection of the writing and the checklist. As many participants forgot to include the total time on one or more checklist submissions, this data was not included in the final analysis. The final essay (Essay 4, Persuasion essay) did not have a checklist for either group. This was done to see if there was any carryover of prior generation treatments to subsequent writing tasks. Additionally, essays were progressively more challenging, starting with Essay 1 as the least challenging to Essay 4 as the most challenging. This increased

[3] Prior to the start of the study, participants who did not submit an assignment / essay were to receive a score of 0. However, all the participants who failed to complete an essay, failed to complete multiple essays. This led to omitting their data from the analysis; therefore, 0 was not used in the scoring of essays.

challenge was from increased length requirements, citation and referencing needs, and paraphrasing requirements in Essays 3 and 4.

6.4 Results

Results looked at three areas separately (Formatting, GSPC, and Content). Linear mixed effects models were used to consider the difference between generation and control conditions over time (Essays 1-4).

6.4.1 Results in Aggregate for Formatting Scores

The overall descriptive results for Formatting scores are shown in Table 6.2.

Table 6.2: Mean Formatting Scores by Treatment for Writing Tasks with Standard Deviation (SD) and 95% Confidence Intervals (CIs).

	Control Condition				**Generation Condition**			
Writing Task	**Essay 1**	**Essay 2**	**Essay 3**	**Essay 4**	**Essay 1**	**Essay 2**	**Essay 3**	**Essay 4**
Mean	1.86	1.93	1.93	2.14	2.24	2.59	2.29	2.35
SD	0.54	0.73	0.83	0.77	0.66	0.62	0.47	0.49
95% CIs	1.55, 2.17	1.51, 2.35	1.45, 2.41	1.70, 2.59	1.89, 2.58	2.27, 2.91	2.05, 2.54	2.10, 2.61

An initial model included fixed effects of condition, treatment coded with control as the reference level, and essay number, with 1 as the baseline. Random intercepts for subject and by-subject random slopes for the effect of treatment condition and essay were included.[4] The Formatting scores showed an overall effect of generation treatment ($\beta = 0.40$, $t = 2.62$, p

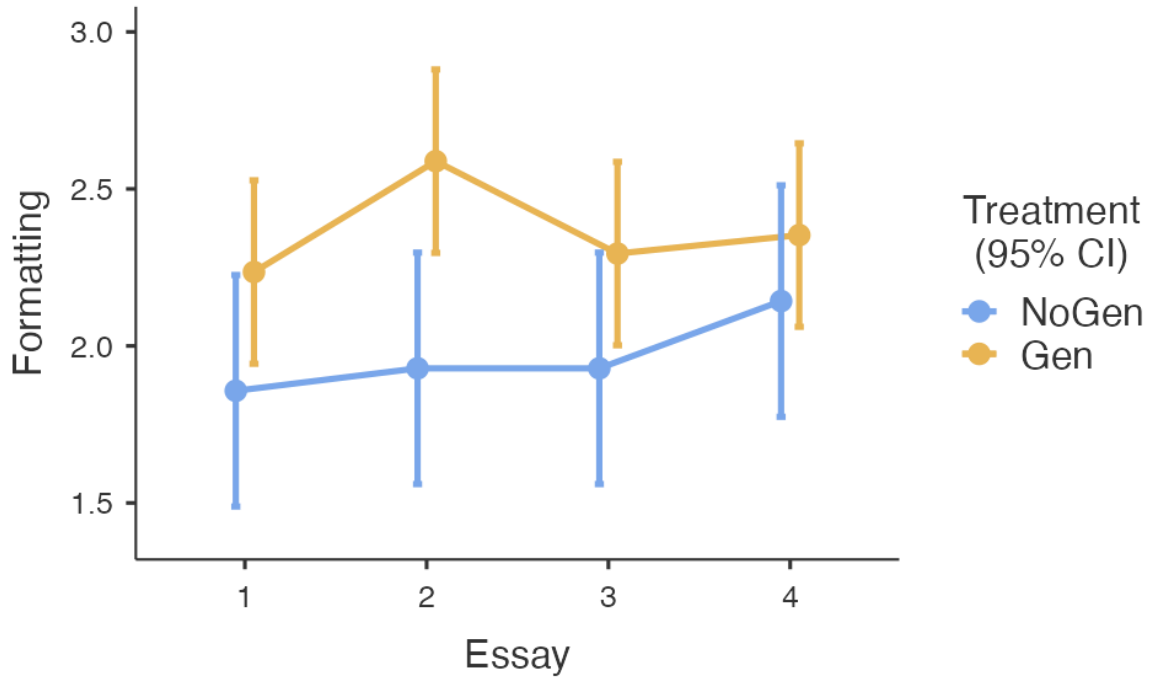
[4] As in study2, there were no "items" in the same way as in the vocabulary study, so these were not included in the random effects structure for this analysis.

= .015), see Table 6.3. Pairwise comparison using Bonferroni adjustment for multiple comparisons showed no differences between conditions for Essay 1 (t = -1.61, p = 1.0), Essay 2 (t = -2.81, p = .17), and Essay 3 (t = -1.56, p = 1.0) or essay 4 (t = -.90, p = 1.0) (see Post Hoc Appendix 6c) There were no significant differences found in Essays or Essay*treatment interactions.

Table 6.3: Model Output for the Effect of Treatment (Reference Level = Control) for Formatting Scores, z and p-values with Control and Confidence Intervals

				95% Confidence Interval				
Names	**Effect**	**Estimate**	**SE**	**Lower**	**Upper**	**df**	**t**	**p**
(Intercept)	(Intercept)	2.1660	0.0771	2.0148	2.317	23.5	28.0842	< .001
Treatment1	Gen - NoGen	0.4034	0.1542	0.1010	0.706	23.5	2.6150	0.015 *
Essay2	2 - 1	0.2122	0.1445	-0.0710	0.495	87.0	1.4686	0.146
Essay3	3 - 1	0.0651	0.1445	-0.2181	0.348	87.0	0.4508	0.653
Essay4	4 - 1	0.2017	0.1445	-0.0815	0.485	87.0	1.3959	0.166
Treatment1 ✻ Essay1	Gen - NoGen ✻ 2 - 1	0.2815	0.2890	-0.2848	0.848	87.0	0.9742	0.333
Treatment1 ✻ Essay2	Gen - NoGen ✻ 3 - 1	-0.0126	0.2890	-0.5790	0.554	87.0	-0.0436	0.965
Treatment1 ✻ Essay3	Gen - NoGen ✻ 4 - 1	-0.1681	0.2890	-0.7344	0.398	87.0	-0.5816	0.562

Figure 6.3: Side by Side Comparison of Treatments, including 95% Confidence Intervals (CI) for Formatting Scores



6.4.2 *Results in Aggregate for Grammar, Spelling, Punctuation, and Capitalization Scores*

The overall descriptive results for GSPC scores are shown in Table 6.4.

Table 6.4: Mean Grammar, Spelling, Punctuation, and Capitalization (GSPC) Scores by Treatment for Writing Tasks with Standard Deviation (SD) and Confidence Intervals (CIs).

	Control Condition				**Generation Condition**			
Writing Task	**Essay 1**	**Essay 2**	**Essay 3**	**Essay 4**	**Essay 1**	**Essay 2**	**Essay 3**	**Essay 4**
Mean	1.86	1.93	1.86	1.93	2.47	2.29	2.41	2.59
SD	0.363	0.475	0.535	0.267	0.624	0.588	0.507	0.507
95% CIs	1.65, 2.07	1.65, 2.20	1.55, 2.17	1.77, 2.08	2.15, 2.79	1.99, 2.60	2.15, 2.67	2.33, 2.85

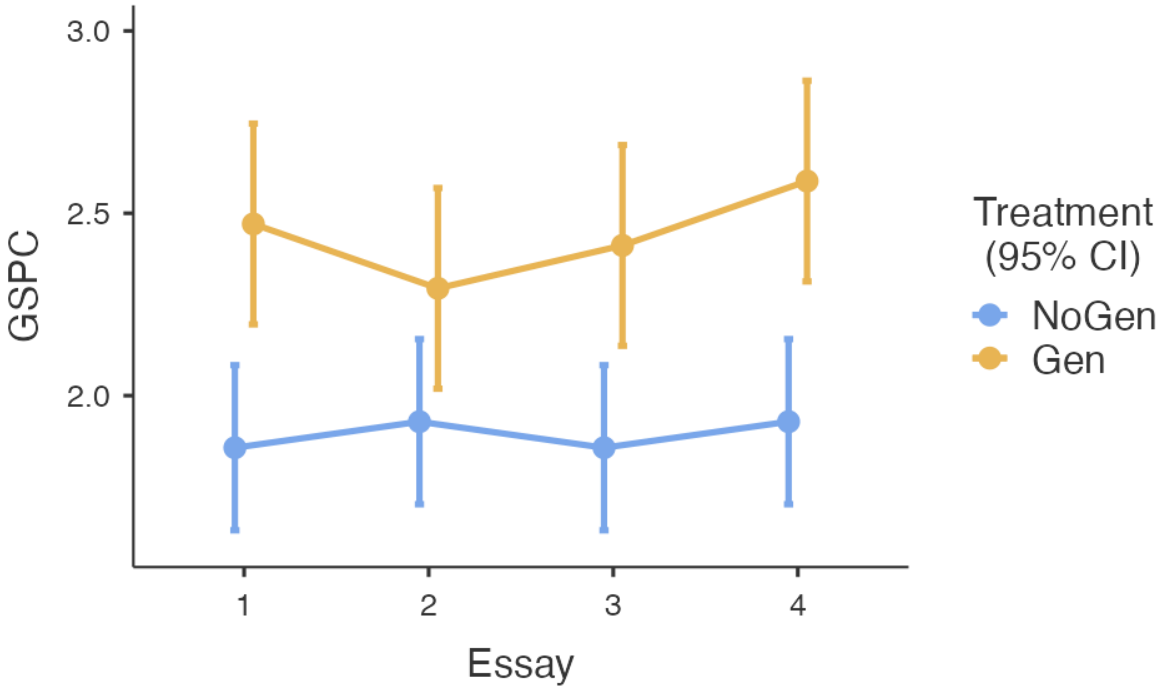
The model used the same fixed and random effects as for the Formatting analysis. The GSPC scores showed an overall effect of generation ($\beta = 0.55$, $t = 4.15$, $p < .001$), see Table

6.5. Figure 6.4 provides a side-by-side comparison of the treatments` GSPC scores and shows a clear effect of treatment where generation illustrates a consistent benefit over time. The greatest difference is seen in Essay 1 and the final essay, Essay 4. Pairwise comparison using Bonferroni adjustment for multiple comparisons showed differences between conditions for Essay 1 (t = -3.48, p = .02) and Essay 4 (t = -3.74, p = .01), with Essay 3 (t = -3.14, p = .07) showing a near significant difference. Essay 2 (t = -2.07, p = 1.0) did not show a difference (see Post Hoc Appendix 6d). There were no significant differences found in Essays or Essay*treatment interactions.

Table 6.5: Model Output for the Effect of Treatment (Reference Level = Control) for GSPC Scores, z and p-values with Control and Confidence Intervals

				95% Confidence Interval				
Names	**Effect**	**Estimate**	**SE**	**Lower**	**Upper**	**df**	**t**	**p**
(Intercept)	(Intercept)	2.1670	0.0661	2.0374	2.297	26.7	32.771	< .001
Treatment1	Gen - NoGen	0.5483	0.1323	0.2891	0.808	26.7	4.146	< .001 *
Essay1	2 - 1	-0.0525	0.0954	-0.2395	0.135	87.0	-0.550	0.583
Essay2	3 - 1	-0.0294	0.0954	-0.2164	0.158	87.0	-0.308	0.759
Essay3	4 - 1	0.0945	0.0954	-0.0925	0.282	87.0	0.991	0.325
Treatment1 ✻ Essay1	Gen - NoGen ✻ 2 - 1	-0.2479	0.1908	-0.6220	0.126	87.0	-1.299	0.197
Treatment1 ✻ Essay2	Gen - NoGen ✻ 3 - 1	-0.0588	0.1908	-0.4329	0.315	87.0	-0.308	0.759
Treatment1 ✻ Essay3	Gen - NoGen ✻ 4 - 1	0.0462	0.1908	-0.3278	0.420	87.0	0.242	0.809

Figure 6.4: Side by Side Comparison of Treatments, including 95% Confidence Intervals (CI) for GSPC Scores



6.4.3 *Results in Aggregate for Content Scores*

Table 6.6: Mean Content Scores by Treatment for Writing Tasks with Standard Deviation (SD) and Confidence Intervals (CIs).

	Control Condition				Generation Condition			
Writing Task	**Essay 1**	**Essay 2**	**Essay 3**	**Essay 4**	**Essay 1**	**Essay 2**	**Essay 3**	**Essay 4**
Mean	1.86	2.00	2.21	2.21	2.47	2.53	2.59	2.65
SD	0.363	0.392	0.426	0.579	0.514	0.514	0.507	0.493
95% CIs	1.65, 2.07	1.77, 2.23	1.97, 2.46	1.88, 2.55	2.21, 2.74	2.26, 2.79	2.33, 2.85	2.39, 2.90

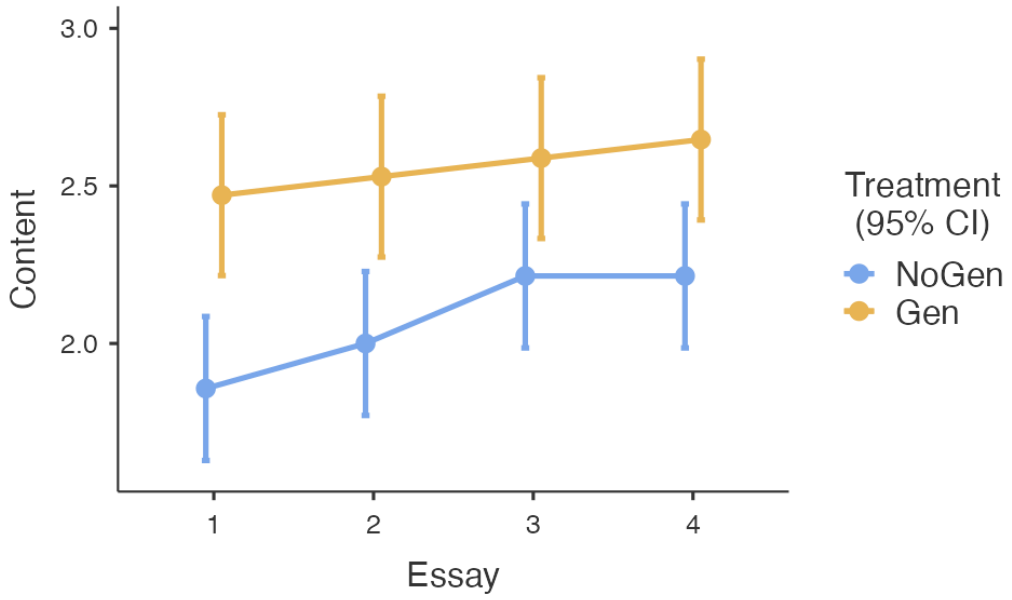
The model used the same fixed and random effects as for the Formatting and GSPC analysis. The Content scores showed a clear effect of generation ($\beta = 0.49$, $t = 4.04$, $p < .001$). Additionally, simple effects were found for Essay 3-1 ($t = 2.41$, $p = .018$) and Essay 4-1 ($t =$

2.71, p = .008), see Table 6.7. There were no significant differences found in Essays 1-2 or any from Essay*treatment interactions. Figure 6.5.1 provides a pairwise comparison of the treatments` by Essay for Content scores using Bonferroni adjustment for multiple comparisons showed differences between conditions for Essay 1 (t = -3.60, p = .02), and a near significant difference for Essay 2 (t = -3.11, p = .07); however, Essay 3 (t = -2.19, p = .87) and Essay 4 (t = -2.54, p = .36) did not show a significant difference in pairwise comparisons, see Post Hoc Appendix 6e. Overall, there was a clear benefit for the generation condition for Content. Although there was also improvement over time for both conditions, generation was higher at all points.

Table 6.7: Model Output for the Effect of Treatment (Reference Level = Control) for Content Scores, z and p-values with Control and Confidence Intervals

			95% Confidence Interval					
Names	**Effect**	**Estimate**	**SE**	**Lower**	**Upper**	**df**	**t**	**p**
(Intercept)	(Intercept)	2.3151	0.0603	2.1970	2.433	27.9	38.404	< .001
Treatment1	Gen - NoGen	0.4874	0.1206	0.2511	0.724	27.9	4.043	< .001 *
Essay1	2 - 1	0.1008	0.0983	-0.0919	0.294	87.0	1.025	0.308
Essay2	3 - 1	0.2374	0.0983	0.0446	0.430	87.0	2.414	0.018 *
Essay3	4 - 1	0.2668	0.0983	0.0741	0.460	87.0	2.713	0.008 *
Treatment1 ✻ Essay1	Gen - NoGen ✻ 2 - 1	-0.0840	0.1967	-0.4695	0.301	87.0	-0.427	0.670
Treatment1 ✻ Essay2	Gen - NoGen ✻ 3 - 1	-0.2395	0.1967	-0.6250	0.146	87.0	-1.218	0.227
Treatment1 ✻ Essay3	Gen - NoGen ✻ 4 - 1	-0.1807	0.1967	-0.5662	0.205	87.0	-0.919	0.361

Figure 6.5: Side by Side Comparison of Treatments, including 95% Confidence Intervals (CI) for Content Scores



6.6 Discussion

This present study aimed to investigate the effectiveness of using generation tasks through post-writing reflective checklists in teaching academic writing to ELLs at universities in Japan. The research questions investigated were: 1) Does the use of generation in post-writing reflective checklists enhance academic writing in the L2 context? 2) Are there differences in efficacy for generation tasks focused on content versus editing (Formatting and GSPC)? 3) Is there a carryover effect of prior generation use on subsequent writing tasks?

The results showed that participants who received the generation treatment had significantly higher scores on all three aspects (Formatting, GSPC, and Content) than the control group. These results indicate that using generation in post-writing checklists may be

an effective way to enhance L2 academic writing. By requiring learners to reflect by generating self-feedback explanations, generation illustrated in this study that it could help reduce Formatting, GSPC, and Content mistakes. The least cognitively demanding of the three was Formatting. While generation showed an effect on Formatting, this effect difference in the final two essays (Essays 3 & 4) was not as strong as in earlier essays as shown in Figure 6.3. This could have been due to the less complex nature and ease of learning with practice, as illustrated in the narrowing of the effect between treatments in Essays 2 and 3, seen in Figure 6.3. After the basic steps and rules for Formatting were practiced in Essays 1 and 2, there was less need for reflection to complete the aspect properly and perhaps less room for improvement.

The GSPC showed a clear effect of generation; however, what was unique to this aspect was the lack of improvement of the control condition. Figure 6.4 illustrates how the generation treatment positively affected the scores for GSPC in Essays 3 and 4, while the control group saw little change. This suggests that generating reflection may have led to more attention to mistakes than reflecting alone. This follows the findings by Avarzamani and Farahian (2019); while reflection in writing among ELLs was limited, it was effective in writing processes. The specific processes assessed around grammar, spelling, punctuation, and capitalization should not have been difficult to correct if time was taken to notice and get clarification (e.g., from a dictionary, a peer, or the instructor). This suggests a further possible reason for the difference between the treatments as a lack of immediate responsibility in justifying why the control group made the mistake. For novice writers, focusing on completing the task is of primary concern (Bereiter & Scardamalia, 1987). If completion was the immediate goal during the reflective task, a lack of reflection and motivation to make

corrections might account for the lower scores. Conversely, the generation condition was required not only to reflect but show their reflective thoughts in writing. By doing this, they presumably already made the mental corrections, and therefore making corrections on the page took less effort. The knowledge-transforming model of more skilled writers argues that skilled writers analyze problems, reflect on tasks, and rework ideas. While both groups in this study were at the same writing level, the fact that generation seems to have improved the analyzation of problems and motivation of the students to rework their writing actively suggests generation may help move writers from novice (knowledge-telling model) to more skilled (knowledge-transforming).

The results demonstrate that generation may also improve the writing of Content as described. The overall effect in Figure 6.5 shows both sets of participants getting better, but the generation group scores started and stayed higher after each essay without the direction changing over time. Essay Content scores did not improve from Essay 1-2, but there was significant improvement in Content scores from Essays 1-3 and Essays 1-4 from where they started, which suggests that critical thinking in locating gaps in writing skills in early tasks may have enhanced writing ability in later essays. While Content was scored separately from the other two aspects investigated, the fact that those skills also showed a positive effect suggests that generation in reflection may have broad applicability to writing improvement among L2 learners. This improvement could be a result of improved metacognitive control from one essay to the next in this study. Tarricone (2011) argues that when reflection is structured, such as in this study, metacognition is linked to awareness and the practice of writing, which improves metacognitive control. As described in the literature on novice and skilled writers (see Hinkel, 2020; Hyland, 2019), the learner`s writing ability affects what

aspect of writing is focused on (Hyland, 2019; Bereiter & Scardamalia, 1987). In the initial Summary Essay, Content scores were lowest for both treatments. However, as students became more comfortable reflecting on their writing through the structured checklist, they improved their Content writing ability. The major difference between the two is the fact that in the generation condition, students had to justify it in writing, possibly solidifying their understanding and, therefore, improving more than the control group.

While the results of the generation treatment were positive, it is important to note that a separate possible reason was the time on task with the checklist for Essays 1 to 3, but not with Essay 4. The exact time that participants worked on the checklist is unknown; however, it is likely that the generation treatment required much more time, and therefore students likely spent more time and attention reflecting on their papers and looking for specific mistakes. One reason for this may be the effect of the writing process implemented at the beginning of the study and used throughout each essay. The process of assignments included five stages, but students were instructed on how to break each of these stages into individual process steps (see Flower & Hayes, 1981; Hacker, Keener, & Kircher, 2009): brainstorming and outlining using a mentored example essay, initial drafting of the essay, in-class peer review, revision of essay, and checklist task with final editing/corrections. In the final checklist task, both groups were instructed to reflect and make corrections as needed. However, the generation group needed to write out explanations for each checklist point in addition to writing a general reflective paragraph on the task.

The findings in this study are in line with Nielsen`s (2014) self-assessment framework and reflection. Nielson puts forth the argument that specific prompts to reflect on writing through a checklist can engage the student to perceive their overall ideas and content in

addition to specific revision tasks. The results of the present study are similar to Vasu et al. (2018), who found self-assessment checklists for argumentative writing to be effective among ELLs in university at writing more systematically. Vasu et al. had students reflect on their writing and make self-assessments of their mistakes. While self-assessment checklists are not the same as reflective proofreading checklists, the role in prompting students to reflect and give self-feedback (assessment) is similar to the present study. However, this present study extends Nielson`s framework and Vasu et al. by showing that the benefits of generation in written form may not only enhance current writing but may transfer to future writing where a checklist or other form of scaffolding is not used. The next and final chapter, Chapter 7, will discuss the findings of the studies conducted in this thesis looking at generation use in L2 vocabulary, reading comprehension, and writing.

6.6.1 Limitations

As with the other studies in this thesis, this study has limitations that should be considered when interpreting the results. Both sample size and the number of groups investigated were limited in size, which may not represent the general learner population. This limitation was a result of constraints in collecting data from only two academic writing classes, which were assigned by the university where the study was conducted. Second, the participants were lower-intermediate English learners (CEFR B1) at a specific university in Japan, which may not represent learners with higher or lower L2 ability or outside of the Japanese university context. Third, participants were placed in the B1 level by a university-developed English placement test that did not include a writing aspect. Therefore, the B1 level may not be an accurate placement of the participants` writing ability. Next, the setting of the study was that of

an actual classroom where both internal and external factors (e.g., requirements from other class or social club activities) may have affected the attention of students during any given class. While precautions to avoid bias were taken, the instructor was the researcher that scored and analyzed the data, and outside scorers were not used, so this should be considered a limitation of the study and its reproducibility. Finally, the study was exploratory in nature, which limited the amount of specificity of aspects (Formatting, GSPC, and Content) that were used in the research. Nevertheless, the study found useful insights into the use of self-feedback through generation checklists in L2 writing. Future studies should be done looking at L2 student writers from different levels and from different cultural backgrounds.

7.0 DISCUSSION & CONCLUSION

7.1 Introduction

This thesis set out to investigate the role of generation in the L2 learning context, focusing on generation`s effectiveness in three areas of L2 learning: vocabulary recall, reading comprehension, and self-feedback through reflective writing checklists. This chapter will begin by providing a summary of each of the studies, including an overview of the findings, before discussing the limitations of the studies. This is then followed by the specific contributions that this thesis has made to the field of SLA. In the final sections, pedagogical implications of the findings on generation are provided along with future directions of research. This is followed by a concluding section that provides key takeaways from the thesis.

7.2 Summary of Studies and Findings

In Chapter 3 of this thesis, I described a previously conducted online classroom research study from Module 2 consisting of two experiments on Japanese university ELLs investigating the role of generation in L2 vocabulary recall. The first experiment led to methodological changes for better recall assessment. The follow-up experiment found no significant benefit for generation on L2 vocabulary recall. Due to issues with online adherence and the focus on vocabulary definition recall, a further study on L2 vocabulary learning using generation was conducted as Study 1 (Chapter 4) of this thesis.

In Chapter 4, Study 1 expanded on Module 2 by examining vocabulary learning over a longer term (three and fifteen weeks) and participants' perceived vocabulary knowledge. This study was conducted face-to-face for better participant adherence. While the generation

treatment showed some vocabulary learning interactions compared to the control group, it did not significantly outperform the baseline treatment in vocabulary learning or perceived knowledge at any test points. However, unlike similar studies (e.g., Barcroft, 2007; 2009) findings, generation did not show a cost.

Study 2 in Chapter 5 investigated the impact of generation tasks on L2 reading comprehension, considering both L1 and L2 generation. The study explored the effects of altering cognitive load demands using two generation tasks and a control condition. No benefits were found from generation in either language, and L2 generation had a cost compared to the control condition for two of the groups. This aligns with the literature on DDs, suggesting that familiarity (e.g., through blocked practice) can be effective in the short term. The role of reducing the burden on cognitive load could not be determined as the L1 generation condition did not differ significantly compared to the L2 generation condition.

Study 3 in Chapter 6 explored self-reflection in L2 academic writing among ELL university students using generation checklists. It found that generation improved all three aspects of writing (Formatting, GSPC, and Content) across four essays. This suggests that self-reflection through generation tasks can enhance L2 academic writing. The study also found a carryover effect from generation tasks in earlier essays to later ones for the most cognitively demanding aspect, Content. However, improvement in Formatting, the least demanding aspect, plateaued in later essays, suggesting that deeper reflection on basic formatting may not be needed once basic formatting is mastered.

7.2.1 Inconsistent Findings across the Studies

This thesis aimed to look at generation`s efficacy in multiple L2 learning contexts. The vocabulary study described in Chapter 4 found that generation did not show a clear benefit, but it also did not come at a cost, as was found by Barcroft (2009) mentioned above in Section 7.2. One possible reason for the difference was the length of time between post-tests. Barcroft assessed generation at two-day and two-week intervals, while the present study looked at recall from a longer-term perspective as the first post-test was conducted at three weeks and the second post-test at 15 weeks from the generation task and vocabulary lesson. Performance and long-term learning with DDs are not always effective in the short-term but tend to improve longer-term learning (Soderstrom & Bjork, 2015). Though a clear benefit was not seen in the present study, the fact that there were interactions found and there was no cost to implementing generation does suggest a distinct pattern of results compared to Barcroft`s (2009) findings, which may be attributed to the clear differences in post-task assessment time (two weeks vs. 15 weeks). In other words, the shorter post-test time period of two-weeks in Barcroft`s study may not have been long enough to show interactions from generation as were found in Study 1.

The role of generation as a method of consolidation was not specifically addressed in this study. However, a possible reason for the interactions and lack of cost found compared to Barcroft (2009) might be somewhat attributed to consolidation if the participants prior to the study had a surface-level knowledge of the vocabulary item at the time of the pre-test and generation task. In this case, the generation task could have been a consolidation task for them. During the initial design of Study 1, generation as a means of consolidation was not a primary point of interest; however, new research into generation and consolidation is currently being

conducted (see Roelle et al., 2022a; 2022b; Schindler & Richter, 2023) and may provide insight into the inconsistencies between efficacy in findings among L1and L2 contexts.

In the reading comprehension context, Study 2 showed a cost to using L2 generation and no clear benefit over control (rereading) from lessening the burden of cognition by using L1. Unlike the vocabulary study in Chapter 4, the testing role of long-term recall was not investigated, as the aim was to assess the effect on reading comprehension by incorporating generation tasks in both L1 and L2. Nevertheless, the immediate post-tests showed blocked practice in the form of rereading of the text to be more effective than generation in L2, which is consistent with the literature on DDs (Soderstrom & Bjork, 2015). This may be due to the short-term performance benefit found when using blocked practice. Another possible reason for the lack of benefit found is the split-attention effect, causing a loss of concentration from text-embedded generative prompts, which were not an aspect of the vocabulary or writing checklist studies and which may account for the lack of efficacy and the cost found in this study. A further possibility might be that the L2 generation task was simply too hard to be of much benefit; therefore, in this case, an overburden of cognitive load may mean this became an undesirable difficulty.

The results found in vocabulary recall and reading comprehension differed greatly from that of the writing context, where a clear benefit of generation implementation was seen in Study 3. Study 3 was progressive in that each task (essay) allowed the participants to build on what was learned and practiced in the previous essay. This progressive aspect could be a reason why the results differed in efficacy from the other studies. Additionally, in Studies 1 and 2, time was limited in completing the task, which was not the case for Study 3. The pressure to complete the tasks in class during the allotted time in the first two studies potentially limited

the time for the learners to process the content, while the writing checklist task encouraged the learners to spend as much time as needed. Furthermore, the first two studies` tasks took place at a time and place, not of the participants` choosing, which may have affected some students who were not able to perform at their best due to outside factors. These factors could include a lack of sleep the night before, other class assignments or tests due on the day the studies` tasks were conducted, or personal issues among participants the day the task was performed. Conversely, the writing study enabled the participants to complete the tasks at a time and place of their choosing as long as it was completed before the deadline. Any of these three aspects from Study 3, or a combination of them, may account for part or all of the differences in efficacy found between the studies. Finally, returning to the carryover effect, the benefit found in the Content aspect of the essay tasks possibly illustrates that cultivating habits over a period of time that is not unrealistic (e.g., during a single class period), may play a strong role in learning cognitively demanding content, such as the Content aspect of the writing tasks. By allowing the learner to hone their writing skills over multiple tasks (essays) and receive feedback during each task, the ability to transfer their ability to the next task likely improves. This was seen in the progressive nature of Study 3. What is not clear, is whether this carryover will continue to tasks outside the context of the Study.

7.3 General Overall Limitations

The studies conducted in this thesis are not without limitations. Firstly, the scope of this research was confined to L2 learning contexts from Japan at a specific Japanese university, and as such, the findings may not be applicable to other populations or disciplines. The research methodology employed in this study also presents certain limitations. The use of quantitative

methods, while helping to see results with less researcher bias and making the findings more generalizable, may oversimplify the data results. This could lead to missed idiosyncrasies around areas such as participant motivation. The time of day, when participants completed the vocabulary study in Chapter 4 differed greatly (early morning versus late afternoon). Since the number of participants in the early morning was severely limited, a possible effect from when tasks were completed was not investigated and should therefore be considered a limitation that can be explored further in future studies.

The sample size and selection process may also limit the applicability of the findings. The participant groupings and number of participants in all studies were selected based on convenience, as groups were derived from university classes decided by the university where the studies took place. Furthermore, the L2 (English) level of the participants was estimated by large-scale placement tests conducted by the university, which may not reflect actual L2 ability as the level assessment did not include speaking or writing aspects of English. Moreover, cultural and institutional factors may have influenced the participants' responses, limiting how these findings can be generalized to other contexts. A possible example of variability between classes, even when estimated to be at the same L2 level, was seen in Study 2 (Chapter 5), which found the different classes responded differently. This suggests that the true level of the students in a particular class may be important in determining a particular benefit. Students at a lower level of reading ability may benefit more from generation (elaborative feedback) tasks, while those at intermediate and advanced levels may not show the same benefit (Bitchener, 2012; Brown, 2017). For example, students at a lower level may benefit more from the process of making connections between new content and their existing schema by building a foundation than those at a higher level where a solid foundation has already been established (Bitchener,

2012). Therefore, ensuring the proper placement of participants according to L2 skills (e.g., reading comprehension) is necessary to determine the reliability of findings on the efficacy of generation in future studies.

The participants in the studies were first-year university students at an elite university in Japan where students had trained for many years in how to take tests to pass exams. The ability of the students to enter the highly competitive university suggests that these students were skilled at taking tests. This ability may play a role in answering test questions correctly despite a possible lack of L2 ability, which could differ from other learners, affecting the generalizability to other L2 learner populations. Additionally, the university where the studies were conducted skewed the variability of gender as an overwhelming majority of the participants were female. Furthermore, the university is well known for L2 language studying in Japan and therefore attracts students who may be more motivated to study L2, which may not be seen in different university contexts, whether in Japan or other countries. Future research could address these limitations by employing different methodologies, expanding the sample size or scope, or exploring other contexts.

Future research using generation in L2 reading comprehension could look at a longer-term perspective by testing at points of multiple weeks instead of or in addition to the immediate comprehension tests as in Study 2. Furthermore, implementing a progressive reading study that builds upon past learning (as was done in Study 3 on writing), may show different results since the time constraints could be eliminated and the reading material could use a book with chapters as testing points with a cumulative post-test to assess total comprehension. Finally, generation tasks could be varied to investigate differences in efficacy when used for encoding versus consolidation. As a classroom action research project, this study

could be implemented in a semester reading course through in-class graded readers where the instructor could control the study.

7.4 Contributions

This dissertation makes several significant contributions to the field of SLA from both theoretical and practical perspectives. The following describes the contributions from each learning aspect investigated (L2 vocabulary acquisition, reading comprehension, and self-reflection in academic writing through checklists) before providing overall contributions from the thesis.

L2 vocabulary recall using Generation was found by Barcroft (2009) to be ineffective in two-day and one-week assessments. As DDs have been shown to be less effective in short-term learning (see Section 1.3.1), this thesis took the findings by Barcroft further and tested the Generation at the three-week and 15-week periods. The vocabulary studies conducted in this thesis supported an overall lack of effectiveness of Generation at longer periods, further supporting Barcroft`s (2009) findings. However, the interactions and trend showing improvement from pre- to post-tests described in Chapter 4 may be significant with a larger sample size. This is where different populations might come in and may show more benefit for higher/lower proficiency students. Consequently, studying more students and subdividing by proficiency may show a more nuanced picture here.

This thesis contributed to what is known about Generation in L2 reading comprehension by investigating the effectiveness of embedded Generation tasks, which showed both a lack of effectiveness and a cost over the control of reading the text twice. Furthermore, it looked at what, to the best of my knowledge, is the only study on Generation and L1 versus L2 usage in

reading comprehension. The differences in results between the reading tasks and groups suggest responses to reading texts and L1/L2 Generation activities may differ greatly depending on the individual, whether it is a result of their existing schema, individual interest, or L2 level. Additionally, it is not possible to tell whether there may be longer term benefits (e.g., a longer-term post-test might show better recall), which is not evident here with the immediate post-test, which illustrates the well-established early cost of using DDs.

Generation use in academic writing through self-feedback took a more exploratory look at the L2 context as there is a lack of available research to build upon. The relatively consistent benefits found provide a strong basis for future studies to expand on. This was particularly evident in the more demanding aspects of writing (GSPC and Content). The methods and techniques of the checklists can be applied in practical settings to both the instructor-led classroom and individual student writers. The carryover effect seen in Study 3 through the use of reflective checklists provides support for enforced self-reflection through Generation tasks and the good habits that such a task can develop in students. Additionally, future research can build upon the SOI model (see Section 2.3.5), where new material will be produced, as was the case in this study on L2 academic writing.

7.5 Implications

This section looks at the implications of this thesis by discussing the pedagogical and future research directions.

7.5.1 Pedagogy

The amount of time and effort available in the L2 classroom is limited, and as such, effective and efficient teaching and learning methods are necessary to achieve the learner's goals in an acceptable time period. Furthermore, students must choose what learning methods to implement, which inherently comes at the cost of excluding other study methods. This thesis looked at the use of Generation from multiple L2 contexts and found mixed results depending on what aspect of L2 is being studied. The following will provide examples of pedagogical applications in the L2 classroom that instructors and students may take into account to maximize the learning of L2 with regard to Generation use.

L2 vocabulary learning is a key part of learning ~~a~~ second language (Schmitt & Schmitt, 2020), and therefore, there is a need to spend focused time learning new terms. Barcroft (2007; 2009; 2015) showed a cost of Generation as the primary DD in learning L2 vocabulary, which was not found in this thesis. However, a clear benefit could not be shown either and the fact that Generation tasks require more effort and time from the learner, suggests that Generation in vocabulary learning may not be an optimal strategy. Therefore, DD use in learning new L2 vocabulary items may show more benefit from other DDs, such as spacing, interleaving, and retrieval, for the initial learning tasks before making the choice whether or not to implement Generation as a method for consolidating learned content. Consequently, Generation alone may not be effective, and other methods in conjunction or in place of it are likely required.

The benefits of Generation in L1 contexts may not be applicable to L2 contexts early in the encoding process. Some CLT proponents (e.g., Chen et al., 2015; 2016; 2018) found that more complexity inhibits the encoding of material to long-term memory storage. As Generation requires more cognitive resources (see Section 2.4.3), less cognitively demanding DDs (e.g.,

Retrieval) may be a better strategy for the L2 learner initially before a suitable schema has been built. Recent research has raised this specific question and is currently being investigated. Findings published after the studies conducted in this thesis look at using Generation with Retrieval practice as a way to complement each other, with Generation used primarily following initial learning (Roelle et al., 2022a; 2022b; Schindler & Richter, 2023) with the purpose of consolidating material already encoded in long-term memory. Generation may have some benefits in consolidating learned vocabulary, but this was beyond the scope of this thesis and, therefore, not investigated. Its use as an initial step in learning L2 vocabulary is not strongly supported and comes with a clear time and effort cost without the appropriate corresponding benefit (Dunlosky, 2013).

A specific strategy that may be implemented would be to provide clear examples of the target term that is related to a subsequent reading context for initial exposure. Following the contextual reading with the vocabulary, Retrieval practice (e.g., flashcards), incorporating spacing and interleaving could be used until the term has been encoded in long-term memory by the learner. It is at this point that the learner may be able to confidently generate new knowledge using the L2 vocabulary in transferring what has been learned to new learning or contexts. By implementing the DDs described here in order, the complexity would likely be reduced, leading to less demand on working memory, hence, less demand on cognitive load during the learning process. In the vocabulary studies in this thesis, the less complex DDs were skipped over for the purpose of implementing Generation, which is a more complex and demanding strategy, at the onset of the learning process, with the hypothesis that it may enhance longer-term recall. This enhancement did not occur; consequently, less demanding or complex strategies during initial exposure to target L2 vocabulary may show more efficacy.

Another context where Generation may not be effective in initial encoding is in L2 reading comprehension. The costs shown by Generation tasks on L2 readers from Study 2, illustrate a difference from similar Generation tasks used in L1 contexts where novel Generation tasks were found to be more effective in reading comprehension than simple tasks without the use of Generation (e.g., Hagen, Braasch, & Braten, 2014; Wood, 1986). Consequently, the extra difficulty incurred from reading in L2 should lead the instructor to provide more opportunities to reread texts for the L2 contexts, which differs from what is argued for the effectiveness of L1 reading strategies (Brown et al., 2014; Dunlosky et al., 2013). This highlights the difference between L1 and L2 learning contexts around reading comprehension. The difference in vocabulary knowledge between L1 and L2 learners may be a reason for this discrepancy in effectiveness. Therefore, providing the L2 learner the opportunity to review or pre-learn difficult vocabulary before the reading activity, as described earlier in this section, would likely benefit L2 readers.

Another possible reason for the lack of effectiveness of Generation found for reading in Study 2 is the immediacy of the assessment. The difference between performance and long-term retention discussed in Section 1.3.1 has been shown in reading tasks where Generation through short-answer questions showed efficacy over simple multiple-choice questions in the longer term; however, in immediate assessment, simplified retrieval through multiple choice questions may be optimal (McDaniel et al., 2007). By initially incorporating comprehension assessment with non-generative multiple-choice retrieval questions and then following the task up with generative short answers questions, a schema can be developed, possibly improving the learning strategy by encoding content with retrieval and consolidating it after initial encoding with generative tasks. The current state in the literature of using retrieval and generative

methods together is still unclear (Brod, 2021; Roelle et al., 2023); however, there are recent studies, though not in an L2 context, that show some promising initial results using Generation as a consolidating method (e.g., Spens & Burgess, 2023; Waldeyer et al., 2020; Wang, Cheng, & Meyer, 2023).

Study 3 found a clear benefit of using Generation in self-feedback in academic writing. These findings support implementing reflective activities to help the learner locate writing mistakes and identify gaps in ability or knowledge. A key difference between using Generation tasks in the vocabulary and reading comprehension studies with that of the academic writing study is the time allotted for the learner to reflect. The studies where Generation was found to be ineffective required the learner to complete a task during a set amount of time. Alternatively, the self-reflective writing task did not limit the students to a single class period or set amount of time and encouraged the learner to take a slower approach to the learning task. Additionally, the requirement of holding content in working memory during the encoding process in Studies 1 and 2, was not applicable to the writing study. This study enabled the students to build up from one essay to the next and establish a habit of paying attention to the task at hand, which may have encouraged more self-reflection as was likely seen in the carryover effect to Essay 4 from Chapter 6 (Study 3), where the checklists were not part of the task.

7.5.2 Future Research Directions

This thesis has investigated the use of Generation in the L2 classroom from multiple contexts within the broader field of SLA. While the findings have contributed valuable insights, there remain several unexplored avenues and potential extensions for future research. This

section offers recommendations for future research to advance our understanding of Generation among L2 contexts.

One promising direction for future research is to expand the scope of the present studies. By studying populations outside of Japanese universities and among non-Japanese students, the role of Generation in learning a language can be expanded and compared. Furthermore, the sample could be grouped so that there is better control over L2 ability, as the present studies used predetermined classes grouped by level according to a university English placement test. The number of participants should be increased with a more balanced gender ratio so that gender can be assessed as a possible factor throughout the studies. As discussed in Section 4.4, gender is important to investigate in the Japanese context due to the traditional marginalization of women in Japanese society and the increased professional opportunities for women with strong English language ability. A broader perspective would allow researchers to better understand the nuances and complexities of Generation among a more diverse population. The methodology used in the current study leaned heavily towards quantitative methods. Future research could benefit from a mixed-methods approach to gain a more comprehensive understanding of Generation in a variety of L2 contexts. Combining both quantitative and qualitative data can provide a richer and more nuanced perspective.

7.6 Conclusion

Given the importance of using effective study methods in learning a language, this thesis investigated the role and efficacy of Generation on L2 learning from three areas of focus: L2 vocabulary, reading comprehension, and self-feedback (reflection) in academic writing.

This chapter concluded the thesis by highlighting the key findings as they related to the initial aims laid out in the introduction chapter. These findings demonstrated that the efficacy of Generation in the L2 context differs from findings in L1 contexts and may even differ within the L2 context itself depending on the type of task (e.g., reading comprehension versus writing). While there is much remaining to be investigated regarding the use of Generation in L2 learning, the studies conducted in this thesis have provided further arguments against using Generation as an initial DD in vocabulary learning, provided evidence of a lack of efficacy in Generation for L2 reading comprehension, and illustrated benefit when Generation is used as self-feedback in L2 writing.

References

Abel, R., & Hänze, M. (2019). Generating causal relations in scientific texts: The long-term advantages of successful generation. *Frontiers in Psychology*, *10*, 199. https://doi.org/doi.org/10.3389/fpsyg.2019.00199

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research*, *87*(3), 659–701.

Agarwal, P. K., & Bain, P. M. (2019). *Powerful teaching: Unleash the science of learning*. John Wiley & Sons.

Agarwal, P. K., Nunes, L. D., & Blunt, J. R. (2021). Retrieval practice consistently benefits student learning: A systematic review of applied research in schools and classrooms. *Educational Psychology Review*, *33*(4), 1409–1453. https://doi.org/doi.org/10.1007/s10648-021-09595-9

Al-Shehri, S., & Gitsaki, C. (2010). Online reading: A preliminary study of the impact of integrated and split-attention formats on L2 students' cognitive load. *ReCALL*, *22*(3), 356–375.

Allen, D., & Nagatomo, D. H. (2019). Investigating the consequential validity of TEAP: Washback to high school learners of English. *Eiken Bulletin*.

Annis, L. F. (1985). Student-generated paragraph summaries and the information-processing theory of prose learning. *The Journal of Experimental Education*, *54*(1), 4–10. https://doi.org/doi.org/10.1080/00220973.1985.10806390

Avarzamani, F., & Farahian, M. (2019). An investigation into EFL learners' reflection in writing and the inhibitors to their reflection. *Cogent Psychology*, *6*(1), 1690817.

Baddeley, A., Eysenck, M. W., & Anderson, M. C. (2014). Memory (2nd ed). *Hoboken: Taylor and Francis*.

Barcroft, J. (2006). Can writing a new word detract from learning it? More negative effects of forced output during vocabulary learning. *Second Language Research*, *22*(4), 487–497.

Barcroft, J. (2007). Effects of Opportunities for Word Retrieval During Second Language Vocabulary Learning. *Language Learning*, *57*(1), 35–56. https://doi.org/10.1111/j.1467-9922.2007.00398.x

Barcroft, J. (2009). Effects of Synonym Generation on Incidental and Intentional L2 Vocabulary Learning During Reading. *TESOL Quarterly*, *43*(1), 79–103. https://doi.org/10.1002/j.1545-7249.2009.tb00228.x

Barcroft, J. (2015). Can Retrieval Opportunities Increase Vocabulary Learning During Reading? *Foreign Language Annals*, *48*(2), 236–249. https://doi.org/10.1111/flan.12139

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Bates, B. (2019). Learning theories simplified: And how to apply them to teaching. *Learning Theories Simplified*, 1–384.

Begg, I., Duft, S., Lalonde, P., Melnick, R., & Sanvito, J. (1989). Memory predictions are based on ease of processing. *Journal of Memory and Language*, *28*(5), 610–632.

Bereiter, C., & Scardamalia, M. (1987). An attainable version of high literacy: Approaches to teaching higher-order skills in reading and writing. *Curriculum Inquiry*, *17*(1), 9–30.

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition*, *35*(2), 201–210. https://doi.org/10.3758/BF03193441

Bird, S. (2011). Effects of distributed practice on the acquisition of second language English syntax—ERRATUM. *Applied Psycholinguistics*, *32*(2), 435–452. https://doi.org/doi:10.1017/S0142716410000172

Bitchener, J. (2012). Written corrective feedback for L2 development: Current knowledge and future research. *Tesol Quarterly*, *46*(4), 855–860.

Bjork, R. A. (1994). Memory and metamemory considerations in the. *Metacognition: Knowing about knowing*, *185*(7.2).

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society*, *2*(59–68).

BjorK & Kroll. (2015). Desirable Difficulties in Vocabulary Learning. *The American Journal of Psychology*, *128*(2), 241. https://doi.org/10.5406/amerjpsyc.128.2.0241

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. *From Learning Processes to Cognitive Processes: Essays in Honor of William K. Estes*, *2*, 35–67.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology*, *64*, 417–444.

Brod, G. (2021). Generative learning: Which strategies for what age? *Educational Psychology Review*, *33*(4), 1295–1318. https://doi.org/10.1007/s10648-020-09571-9

Bown, A. (2017). Elaborative Feedback to Enhance Online Second Language Reading Comprehension. *English Language Teaching*, *10*(12), 164–171.

Brown, D. (2012). The written corrective feedback debate: Next steps for classroom teachers and practitioners. *TESOl Quarterly*, *46*(4), 861–867.

Brown, J. D. (1997). Testing washback in language education. *PASAA Journal*, *27*, 64–79.

Brown, P. C., Roediger III, H. L., & McDaniel, M. A. (2014). *Make it stick: The science of successful learning.*

Bruen, J., & Kelly, N. (2017). Using a shared L1 to reduce cognitive overload and anxiety levels in the L2 classroom. *The Language Learning Journal*, *45*(3), 368–381. https://doi.org/doi.org/10.1080/09571736.2014.908405

Brunmair, M., & Richter, T. (2019). Similarity matters: A meta-analysis of interleaved learning and its moderators. *Psychological Bulletin*, *145*(11), 1029–1052. https://doi.org/10.1037/bul0000209

Cadaret, C. N., & Yates, D. T. (2021). Homework questions designed to require higher-order cognitive skills in an undergraduate animal physiology course did not produce desirable difficulties, testing effects, or improvements in information retention. *Journal of Animal Science*, *99*(9), skab246.

Callender, A. A., & McDaniel, M. A. (2009). The limited benefits of rereading educational texts. *Contemporary Educational Psychology*, *34*(1), 30–41. https://doi.org/10.1016/j.cedpsych.2008.07.001

Camp, R. (1998). Portfolio reflection: The basis for dialogue. *The Clearing House*, *72*(1), 10–12. https://doi.org/10.1080/00098659809599377

Carpenter, S. K., & Mueller, F. E. (2013). The effects of interleaving versus blocking on foreign language pronunciation learning. *Memory & Cognition*, *41*(5), 671–682.

Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory & Cognition*, *36*(2), 438–448.

Carvalho, P. F., & Goldstone, R. L. (2015). What you learn is more than what you see: What can sequencing effects tell us about inductive category learning? *Frontiers in Psychology*, *6*, 505. https://doi.org/doi.org/10.3389/fpsyg.2015.00505

Carvalho, P. F., & Goldstone, R. L. (2017). The sequence of study changes what information is attended to, encoded, and remembered during category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(11), 1699.

Cepeda, N. J., Coburn, N., Rohrer, D., Wixted, J. T., Mozer, M. C., & Pashler, H. (2009). Optimizing distributed practice: Theoretical analysis and practical implications. *Experimental Psychology*, *56*(4), 236–246. https://doi.org/10.1027/1618-3169.56.4.236

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science*, *19*(11), 1095–1102.

Chen, O., Castro-Alonso, J. C., Paas, F., & Sweller, J. (2018). Undesirable difficulty effects in the learning of high-element interactivity materials. *Frontiers in Psychology*, 1483.

Chen, O., Kalyuga, S., & Sweller, J. (2015). The worked example effect, the generation effect, and element interactivity. *Journal of Educational Psychology*, *107*(3), 689.

Chen, O., Kalyuga, S., & Sweller, J. (2016). Relations between the worked example and generation effects on immediate and delayed tests. *Learning and Instruction*, *45*, 20–30.

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, *13*(2), 145–182. https://doi.org/10.1016/0364-0213(89)90002-5

Clarke, T., Ayres, P., & Sweller, J. (2005). The impact of sequencing and prior knowledge on learning mathematics through spreadsheet applications. *Educational Technology Research and Development*, *53*(3), 15–24.

Cobb, T. (1999). Breadth and depth of lexical acquisition with hands-on concordancing. *Computer Assisted Language Learning*, *12*(4), 345-360

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213–238.

Daise, D., & Norloff, C. (2019). *Q: Skills for Success: Level 4: Reading and Writing Student Book with iQ Online Practice* (Third). Oxford University Press.

DeWinstanley, P. A., & Bjork, E. L. (2004). Processing strategies and the generation effect: Implications for making a better reader. *Memory & Cognition*, *32*(6), 945–955.

Didau, D. (2015). *What if everything you knew about education was wrong?* Crown House Publishing.

Diemand-Yauman, C., Oppenheimer, D. M., & Vaughan, E. B. (2010). Fortune favors the (): Effects of disfluency on educational outcomes. *Cognition*, *118*(1), 111–115. https://doi.org/10.1016/j.cognition.2010.09.012

Doctorow, M., Wittrock, M. C., & Marks, C. (1978). Generative processes in reading comprehension. *Journal of Educational Psychology*, *70*(2), 109.

Dong, A., Jong, M. S.-Y., & King, R. B. (2020). How does prior knowledge influence learning engagement? The mediating roles of cognitive load and help-seeking. *Frontiers in Psychology*, *11*, 591203. https://doi.org/10.3389/fpsyg.2020.591203

Dunlosky, J. (2013). Strengthening the student toolbox: Study strategies to boost learning. *American Educator*, *37*(3), 12–21.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*(1), 4–58. https://doi.org/10.1177/1529100612453266

Dykes, R., & Hauca, M. (2019). *Sans Forgetica: Typography's effect on ESL/EFL reading comprehension*. 53–60. https://koreatesol.org/sites/default/files/pdf_publications/KOTESOL.Proceedings.2019.pdf#page=63

EF English Proficiency Index (EPI). (2022). *The world's largest ranking of countries and regions by English skills*. https://www.ef.com/wwen/epi/

Elgort, I. (2011). Deliberate learning and vocabulary acquisition in a second language. *Language Learning*, *61*(2), 367–413. https://doi.org/10.1111/j.1467-9922.2010.00613.x

Elgort, I., & Nation, P. (2010). Vocabulary learning in a second language: Familiar answers to new questions. In *Conceptualising 'learning'in applied linguistics* (pp. 89-104). London: Palgrave Macmillan UK.

Ellis, N. C. (1994). *Consciousness in second language learning: Psychological perspectives on the role of conscious processes in vocabulary acquisition.*

Ellis, R. (1997). *SLA research and language teaching.* ERIC.

Enser, Z., & Enser, M. (2020). *Fiorella & Mayer's generative learning in action*. John Catt Educational.

Evans, M. S. (2011). Reading bilinguals reading: First language use and comprehension monitoring in the reading of different textual genres. *New Zealand Studies in Applied Linguistics*, *17*(2), 53–69.

Evans, N. W., Hartshorn, K. J., McCollum, R. M., & Wolfersberger, M. (2010). Contextualizing corrective feedback in second language writing pedagogy. *Language Teaching Research*, *14*(4), 445–463.

Farahian, M., Avarzamani, F., & Rajabi, Y. (2021). Reflective thinking in an EFL writing course: To what level do portfolios improve reflection in writing? *Thinking Skills and Creativity, 39*, 100759. https://doi.org/10.1016/j.tsc.2020.100759

Finn, B., Thomas, R., & Rawson, K. A. (2018). Learning more from feedback: Elaborating feedback with examples enhances concept learning. *Learning and Instruction*, *54*, 104–113.

Fiorella, L., & Mayer, R. E. (2015). *Learning as a generative activity*. Cambridge university press.

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication*, *32*(4), 365–387.

Flowers, B. S. (1981). Madman, architect, carpenter, judge: Roles and the writing process. *Language Arts*, *58*(7), 834–836.

Gallucci, M. (2019). *GAMLj: General analyses for linear models.[jamovi module]*.

Garnett, S. (2020). *Cognitive load theory: A handbook for teachers*. Crown House Publishing Ltd.

Gawande, A. (2011). The Checklist Manifesto: How to Get Things Right (Reprint edition.). *New York: Picador*.

Geary, D. C. (2008). An evolutionarily informed education science. *Educational Psychologist*, *43*(4), 179–195.

Graham, S., & Sandmel, K. (2011). The process writing approach: A meta-analysis. *The Journal of Educational Research*, *104*(6), 396–407.

Gu, P. Y. (2003). Vocabulary learning in a second language: Person, task, context and strategies. *Tesl-Ej*, *7*(2), 1–25.

Hacker, D. J., Keener, M. C., & Kircher, J. C. (2009). Writing is applied metacognition. *Handbook of Metacognition in Education*, 166–184.

Hagen, Å. M., Braasch, J. L., & Bråten, I. (2014). Relationships between spontaneous note-taking, self-reported strategies and comprehension when reading multiple texts in different task conditions. *Journal of Research in Reading*, *37*(S1), S141–S157. https://doi.org/10.1111/j.1467-9817.2012.01536.x

Hamada, M., & Koda, K. (2008). Influence of first language orthographic experience on second language decoding and word learning. *Language Learning*, *58*(1), 1–31.

Harmer, J. (2007). The practice of English language teaching 4th ed. *England: Pearson Education Limited*.

Harrington, M., & Sawyer, M. (1992). L2 working memory capacity and L2 reading skill. *Studies in Second Language Acquisition*, *14*(1), 25–38.

Harris, A. (2018). *Sans Forgetica: New typeface designed to help students study*. https://www.rmit.edu.au/news/all-news/2018/oct/sans-forgetica-news-story

Hattie, J. (2012). *Visible learning for teachers: Maximizing impact on learning*. Routledge.

Hausman, H., & Kornell, N. (2014). Mixing topics while studying does not enhance learning. *Journal of Applied Research in Memory and Cognition*, *3*(3), 153–160. https://doi.org/10.1016/j.jarmac.2014.03.003

Hill, M., & Laufer, B. (2003). *Type of task, time-on-task and electronic dictionaries in incidental vocabulary acquisition*.

Hinkel, E. (2020). *Teaching academic L2 writing: Practical techniques in vocabulary and grammar*. Routledge.

Hulstijn, J. H., & Laufer, B. (2001). Some empirical evidence for the involvement load hypothesis in vocabulary acquisition. *Language Learning*, *51*(3), 539–558. https://doi-org.bham-ezproxy.idm.oclc.org/10.1111/0023-8333.00164Citations: 295

Hung, H. C. M. (2009). *Applying cognitive load theory in reading comprehension*. 184–196. https://camtesol.org/Download/Earlier_Publications/Selected_Papers_Vol.5_2009.pdf#page=196

Hung, H.-T. (2015). Intentional vocabulary learning using digital flashcards. *English Language Teaching*, *8*(10), 107–112.

Hyland, K. (2019). *Second language writing*. Cambridge university press.

Hyland, K., & Hyland, F. (2019). *Feedback in Second Language Writing*. Cambridge university press.

Iwaniec, J. (2019). Language learning motivation and gender: The case of Poland. *International Journal of Applied Linguistics*, *29*(1), 130–143.

Iwata, A. (2012). The Effectiveness of Summary-Writing Activities on the Improvement of Japanese High School EFL Students' Writing Abilities. *Research Bulletin of English Teaching*, 1–22.

Jacoby, J., Chestnut, R. W., & Fisher, W. A. (1978). A behavioral process approach to information acquisition in nondurable purchasing. *Journal of Marketing Research*, *15*(4), 532–544.

Jagaiah, T., Howard, D., & Olinghouse, N. (2019). Writer's checklist: A procedural support for struggling writers. *The Reading Teacher*, *73*(1), 103–110. https://doi.org/doi.org/10.1002/trtr.1802

James, M. A. (2006). Transfer of learning from a university content-based EAP course. *Tesol Quarterly*, *40*(4), 783-806.

Kalyuga, S. (2010). *Schema acquisition and sources of cognitive load.* https://doi.org/10.1017/CBO9780511844744.005

Karpicke, J. D. (2016). A powerful way to improve learning and memory. *Psychological Science Agenda*, *30*(6).

Karpicke, J. D., & Bauernschmidt, A. (2011). Spaced retrieval: Absolute spacing enhances learning regardless of relative spacing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(5), 1250.

Karpicke, J. D., Butler, A. C., & Roediger III, H. L. (2009). Metacognitive strategies in student learning: Do students practise retrieval when they study on their own? *Memory*, *17*(4), 471–479.

Karpicke, J. D., & Grimaldi, P. J. (2012). Retrieval-based learning: A perspective for enhancing meaningful learning. *Educational Psychology Review*, *24*(3), 401–418.

Karpicke, J. D., & Roediger III, H. L. (2008). The critical importance of retrieval for learning. *Science*, *319*(5865), 966–968.

Katzir, T., Hershko, S., & Halamish, V. (2013). The effect of font size on reading comprehension on second and fifth grade children: Bigger is not always better. *PloS One*, *8*(9), e74061.

Kirkland, M. R., & Saunders, M. A. P. (1991). Maximizing student performance in summary writing: Managing cognitive load. *Tesol Quarterly*, *25*(1), 105–121.

Kirschner, P. A., Ayres, P., & Chandler, P. (2011). Contemporary cognitive load theory research: The good, the bad and the ugly. *Computers in Human Behavior*, *27*(1), 99–105.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, *41*(2), 75–86.

Kirschner, P., & Hendrick, C. (2020). *How learning happens: Seminal works in educational psychology and what they mean in practice*. Routledge.

Kobayashi, Y. (2002). The role of gender in foreign language learning attitudes: Japanese female students' attitudes towards English learning. *Gender and Education*, *14*(2), 181–197.

Kohn, A. (1993). The trouble with gold stars, incentive plans, A's, praise and other bribes. *Alfie Kohn. Houghtin Mufflin, Boston*.

Kohn, A. (2011). The case against grades. *Educational Leadership*, *69*(3), 28–33.

Kohn, A. (2013). The case against grades. *Counterpoints*, *451*, 143–153.

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review*, *14*(2), 219–224.

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological Science*, *19*(6), 585–592.

Krashen, S. (1989). *We Acquire Vocabulary and Spelling by Reading: Additional Evidence for the Input Hypothesis*. 26.

Krashen, S. D. (1993). Comments on Stephen D. Krashen's" teaching issues: Formal grammar instruction". The effect of formal grammar teaching: Still peripheral. *Tesol Quarterly*, *27*(4), 722–725.

Krashen, S. D. (1999). Seeking a role for grammar. A review of some recent studies. *Foreign Language Annals*, *32*(2), 245–254.

Kyun, S., Kalyuga, S., & Sweller, J. (2013). The effect of worked examples when learning to write essays in English literature. *The Journal of Experimental Education*, *81*(3), 385–408.

Lee, I. (2013). Research into practice: Written corrective feedback. *Language Teaching*, *46*(1), 108–119.

Leroy, S. (2009). Why is it so hard to do my work? The challenge of attention residue when switching between work tasks. *Organizational Behavior and Human Decision Processes*, *109*(2), 168–181. https://doi.org/10.1016/j.obhdp.2009.04.002

Li, J., Zhang, E.-H., He, X., Zhang, H., Gou, H., Wang, X., Wang, S., & Cao, H.-W. (2022). Retrieval practice enhances learning and memory retention of French words in Chinese-English bilinguals. *Lingua*, *272*, 103294.

Lotfolahi, A. R., & Salehi, H. (2017). Spacing effects in vocabulary learning: Young EFL learners in focus. *Cogent Education*, *4*(1), 1287391. https://doi.org/10.1080/2331186X.2017.1287391

Lovell, O. (2020). *Sweller's Cognitive Load Theory in Action*. John Catt Educational.

McCurdy, M. P., Viechtbauer, W., Sklenar, A. M., Frankenstein, A. N., & Leshikar, E. D. (2020). Theories of the generation effect and the impact of generation constraint: A meta-analytic review. *Psychonomic Bulletin & Review*, *27*, 1139–1165.

McCutchen, D. (2011). From novice to expert: Implications of language skills and writing-relevant knowledge for memory during the development of writing skill. *Journal of Writing Research*, *3*(1), 51–68. https://doi.org/doi.org/10.17239/jowr-2011.03.01.3

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology*, *103*(2), 399–414. https://doi.org/10.1037/a0021782

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, *19*(4–5), 494–513.

McDaniel, M. A., Waddill, P. J., & Einstein, G. O. (1988). A contextual account of the generation effect: A three-factor theory. *Journal of Memory and Language*, *27*(5), 521–536.

McDaniel, M. A., Wildman, K. M., & Anderson, J. L. (2012). Using quizzes to enhance summative-assessment performance in a web-based class: An experimental study. *Journal of Applied Research in Memory and Cognition*, *1*(1), 18–26. https://doi.org/10.1016/j.jarmac.2011.10.001

M.E.X.T. (2020). *The National Curriculum Standards for Grade 5 and Grade 6 in Elementary School Section 10 Foreign Languages*. Ministry of Education, Culture, Sports, Science and Technology (M.E.X.T.). https://www.mext.go.jp/content/20201218-kyoiku01-000011246_2.pdf

Mihaljević Djigunović, J. (1993). Effects of gender on some learner-dependent variables in foreign language learning. *Studia Romanica et Anglica Zagrabiensia: Revue Publiée Par Les Sections Romane, Italienne et Anglaise de La Faculté Des Lettres de l'Université de Zagreb*, *38*, 169–180.

Montero-SaizAja, A. (2021). Gender-based Differences in EFL Learners' Language Learning Strategies and Productive Vocabulary. *Theory and Practice of Second Language Acquisition*, *7*(2), 83–107.

Murre, J. M., & Dros, J. (2015). Replication and analysis of Ebbinghaus' forgetting curve. *PloS One*, *10*(7), e0120644.

Nakata, T. (2011). Computer-assisted second language vocabulary learning in a paired-associate paradigm: A critical investigation of flashcard software. *Computer Assisted Language Learning*, *24*(1), 17–38.

Nakata, T. (2015). Effects of expanding and equal spacing on second language vocabulary learning: Does gradually increasing spacing increase vocabulary learning? *Studies in Second Language Acquisition*, *37*(4), 677–711.

Nakata, T. (2017). Does repeated practice make perfect? The effects of within-session repeated retrieval on second language vocabulary learning. *Studies in Second Language Acquisition*, *39*(4), 653–679. https://doi.org/doi.org/10.1017/S0272263116000280 [Opens in a new window]

Nakata, T., & Suzuki, Y. (2019). Mixing Grammar Exercises Facilitates Long-Term Retention: Effects of Blocking, Interleaving, and Increasing Practice. *The Modern Language Journal*, modl.12581. https://doi.org/10.1111/modl.12581

Nation, I. (2006). How Large a Vocabulary is Needed For Reading and Listening? *The Canadian Modern Language Review*, *63*(1), 59–82. https://doi.org/10.3138/cmlr.63.1.59

Nation, I. (2009). *Teaching ESL/EFL Reading and Writing*. Routledge.

Nation, I. S. (2001). *Learning vocabulary in another language* (Vol. 10). Cambridge university press Cambridge.

Nation, I. S. P., & Waring, R. (2019). *Teaching extensive reading in another language*. Routledge.

Nation, I. S., & Webb, S. A. (2011). *Researching and analyzing vocabulary*. Heinle, Cengage Learning Boston, MA.

Nation, P. (2008). Lexical awareness in second language learning. *Encyclopedia of Language and Education*, *6*, 167–179.

Nation, P. (2009). *4000 Essential English Words 6*. 2009. Compass publishing.

Nation, P. (2018). *4000 Essential English Words 5 Second Ed.* Compass Publishing.

Nation, P. (2020). The different aspects of vocabulary knowledge. In *The Routledge handbook of vocabulary studies* (pp. 15–29). Routledge.

Nation, P. & Meara, P. (2002). *Vocabulary Vocabulary. In N. Schmitt(Ed.), An introduction to applied linguistics (pp. 35-54).* Oxford University Press.

Nawal, A. F. (2017). Cognitive load theory in the context of second language academic writing. *Higher Education Pedagogies*, *3*(1), 385–402.

Nenotek, S. A., Tlonaen, Z. A., & Manubulu, H. A. (2022a). Exploring university students' difficulties in writing english academic essay. *Al-Ishlah: Jurnal Pendidikan*, *14*(1), 909–920.

Nenotek, S. A., Tlonaen, Z. A., & Manubulu, H. A. (2022b). Exploring university students' difficulties in writing english academic essay. *Al-Ishlah: Jurnal Pendidikan*, *14*(1), 909–920.

Nezhad, A. N., Moghali, M., & Soori, A. (2015). Explicit and implicit learning in vocabulary acquisition. *Asian Journal of Education and E-Learning*, *3*(1), 2321–2454.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, *31*(2), 199–218.

Nielsen, K. (2014). Self-assessment methods in writing instruction: A conceptual framework, successful practices and essential strategies. *Journal of Research in Reading*, *37*(1), 1–16.

Nihalani, P. K., & Robinson, D. H. (2022). Balancing collaboration and cognitive load to optimize individual and group desirable difficulties. *Journal of Educational Computing Research*, *60*(2), 433–454.

Nowbakht, M., & Shahnazari, M. (2015). The comparative effects of comprehensible input, output and corrective feedback on the receptive acquisition of L2 vocabulary items. *Advances in Language and Literary Studies*, *6*(4), 103–114.

Nückles, M., Roelle, J., Glogger-Frey, I., Waldeyer, J., & Renkl, A. (2020). The self-regulation-view in writing-to-learn: Using journal writing to optimize cognitive load in self-regulated learning. *Educational Psychology Review*, *32*, 1089–1126.

Otomo, R., & Danping, W. (2016). English language testing of very young children: The case of Japan. *Cogent Education.*, *3*(1), 1209802.

Paas, F., & van Merriënboer, J. J. (2020). Cognitive-load theory: Methods to manage working memory load in the learning of complex tasks. *Current Directions in Psychological Science*, *29*(4), 394–398. https://doi.org/doi.org/10.1177/0963721420922183

Pan, S. C., Tajran, J., Lovelett, J., Osuna, J., & Rickard, T. C. (2019). Does interleaved practice enhance foreign language learning? The effects of training schedule on Spanish verb conjugation skills. *Journal of Educational Psychology*, *111*(7), 1172.

Pellicer-Sánchez, A., Vilkaitė-Lozdienė, L., & Siyanova-Chanturia, A. (2021). 8 Examining L2 Learners' Confidence of Collocational Knowledge. *Vocabulary Theory, Patterning and Teaching*, *152*, 121.

Persellin, D. C., & Daniels, M. B. (2018). *A concise guide to teaching with desirable difficulties*. Stylus Publishing, LLC.

Peters, E. (2016). The learning burden of collocations: The role of interlexical and intralexical factors. *Language Teaching Research*, *20*(1), 113-138.

Prodromou, L. (1995). *The backwash effect: From testing to teaching*. *ELT Journal*, Volume 49, Issue 1, January 1995, Pages 13–25, https://doi.org/10.1093/elt/49.1.13

R Core Team. (2021). *A Language and environment for statistical computing. 2019. R Foundation for Statistical Computing, Vienna, Austria*.

Richards, J. C., & Rodgers, T. S. (2014). *Approaches and methods in language teaching*. Cambridge university press.

Richland, L. E., Bjork, R. A., Finley, J. R., & Linn, M. C. (2005). *Linking cognitive science to education: Generation and interleaving effects*. *27*(27).

Roberts, R., & Kreuz, R. (2015). *Becoming fluent: How cognitive science can help adults learn a foreign language*. Mit Press.

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–27. https://doi.org/10.1016/j.tics.2010.09.003

Roediger III, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255.

Roelle, J., Froese, L., Krebs, R., Obergassel, N., & Waldeyer, J. (2022a). Sequence matters! Retrieval practice before generative learning is more effective than the reverse order. *Learning and Instruction*, *80*, 101634. https://doi.org/doi.org/10.1016/j.learninstruc.2022.101634

Roelle, J., Schweppe, J., Endres, T., Lachner, A., von Aufschnaiter, C., Renkl, A., Eitel, A., Leutner, D., Rummer, R., & Scheiter, K. (2022b). Combining retrieval practice and generative learning in educational contexts. *Zeitschrift Für Entwicklungspsychologie Und Pädagogische Psychologie*.

Rohman, D. G. (1965). Pre-writing the stage of discovery in the writing process. *College Composition and Communication*, *16*(2), 106–112.

Rohrer, D., Dedrick, R. F., Hartwig, M. K., & Cheung, C.-N. (2020). A randomized controlled trial of interleaved mathematics practice. *Journal of Educational Psychology*, *112*(1), 40.

Rohrer, D., Dedrick, R. F., & Stershic, S. (2015). Interleaved practice improves mathematics learning. *Journal of Educational Psychology*, *107*(3), 900.

Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics problems improves learning. *Instructional Science*, *35*(6), 481–498.

Romney, C. (2019). 31 Desirable Difficulties in Reading via Typographic Disfluency and L2 Learners. *PanSIG Journal*.

Ryan, S. (2008). *THE IDEAL L2 SELVES OF JAPANESE LEARNERS OF ENGLISH*. 300.

Sans Forgetica—Font Family (Typeface) Free Download TTF, OTF - Fontmirror.com. (n.d.). Retrieved April 17, 2023, from https://www.fontmirror.com/sans-forgetica

Schindler, J., & Richter, T. (2023). Text Generation Benefits Learning: A Meta-Analytic Review. *Educational Psychology Review*, *35*(2), 44.

Schmitt, D., Schmitt, N., & Mann, D. (2011). *Mastering the academic word list*. Pearson Longman.

Schmitt, N. (1997). Vocabulary learning strategies. *Vocabulary: Description, Acquisition and Pedagogy*, *199227*.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Springer.

Schmitt, N., & Schmitt, D. (2020). *Vocabulary in language teaching*. Cambridge university press.

Schneider, V. I., Healy, A. F., & Bourne Jr, L. E. (2002). What is learned under difficult conditions is hard to forget: Contextual interference effects in foreign vocabulary acquisition, retention, and transfer. *Journal of Memory and Language*, *46*(2), 419–440.

Schiano, B. (2021, June). Reducing Cognitive Overload While Teaching: Simple Practices to Help Educators Stay Focused in the Classroom. *Harvard Business Publishing Education, June 2021*. https://hbsp.harvard.edu/inspiring-minds/reducing-cognitive-overload-while-teaching

Schunk, D. H. (2012). *Learning theories an educational perspective*. Pearson Education, Inc.

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science.*, *10*(2), 176–199. https://doi.org/10.1177/1745691615569000

Spens, E., & Burgess, N. (2023). A generative model of memory construction and consolidation. *BioRxiv*, 2023–01. https://doi.org/10.1101/2023.01.19.524711

Study Using the Pomodoro Technique. (2021, 11). College of New Caledonia. https://cnc.bc.ca/services/prince-george/academic-success-centre/student-support-advice/pomodoro-technique

Sudjasmara, D. B. (2021). The Backwash Effect of Post-Graduate Study Entrance Test on English Competence of Candidate Academics. *JEPAL (Journal of English Pedagogy and Applied Linguistics)*, *1*(2), 76–87.

Sun, H., & Fang, S. (2022). Paired-Associate Second Language Vocabulary Learning: The Role of L1 Translation Familiarity. *Journal of Asia TEFL*, *19*(1), 50.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285.

Sweller, J. (2011). Cognitive load theory. In *Psychology of learning and motivation* (Vol. 55, pp. 37–76). Elsevier.

Tarricone, P. (2011). *The taxonomy of metacognition*. Psychology Press.

Taylor, A., Sanson, M., Burnell, R., Wade, K. A., & Garry, M. (2020). Disfluent difficulties are not desirable difficulties: The (lack of) effect of Sans Forgetica on memory. *Memory*, *28*(7), 850–857. https://doi.org/doi.org/10.1080/09658211.2020.1758726

Tokuhama-Espinosa, T. (2019). The Learning Sciences Framework in Educational Leadership. *Frontiers in Education*, *4*. https://www.frontiersin.org/articles/10.3389/feduc.2019.00136

Tsao, J.-J., Tseng, W.-T., & Wang, C. (2017). The effects of writing anxiety and motivation on EFL college students' self-evaluative judgments of corrective feedback. *Psychological Reports*, *120*(2), 219–241.

Van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics*, *34*(4), 457–479.

Vandergrift, L. (2007). Recent developments in second and foreign language listening comprehension research. *Language Teaching*, *40*(3), 191–210.

Vasu, K., Nimehchisalem, V., Fung, Y. M., & Rashid, S. M. (2018). The usefulness and effectiveness of argumentative writing self-assessment checklist in undergraduate writing classrooms. *International Journal of Academic Research in Business and Social Sciences*, *8*(4), 202–219.

Walczyk, J. J., & Hall, V. C. (1989). Effects of examples and embedded questions on the accuracy of comprehension self-assessments. *Journal of Educational Psychology*, *81*(3), 435.

Waldeyer, J., Heitmann, S., Moning, J., & Roelle, J. (2020). Can generative learning tasks be optimized by incorporation of retrieval practice? *Journal of Applied Research in Memory and Cognition*, *9*(3), 355–369. https://doi.org/10.1016/j.jarmac.2020.05.001

Wang, Z., Gong, S.-Y., Xu, S., & Hu, X.-E. (2019). Elaborated feedback and learning: Examining cognitive and motivational influences. *Computers & Education*, *136*, 130–140.

Webb, S. (Ed.). (2020). The Routledge Handbook of Vocabulary Studies (1st ed.). Routledge. https://doi.org/10.4324/9780429291586

Webb, S. (2020). Approaches to learning, testing and researching L2 vocabulary. *Approaches to Learning, Testing and Researching L2 Vocabulary*, 1–240.

Webb, S., & Chang, A. C.-S. (2015). How does prior word knowledge affect vocabulary learning progress in an extensive reading program? *Studies in Second Language Acquisition*, *37*(4), 651–675. https://doi.org/10.1017/S0272263114000606[Opens in a new window

Webb, S., & Nation, P. (2017). *How vocabulary is learned*. Oxford University Press.

Webb, S., Uchihara, T., & Yanagisawa, A. (2023). How effective is second language incidental vocabulary learning? A meta-analysis. *Language Teaching*, *56*(2), 161–180.

Webb, S., Yanagisawa, A., & Uchihara, T. (2020). How effective are intentional vocabulary-learning activities? A meta-analysis. *The Modern Language Journal*, *104*(4), 715–738.

Weinstein, Y., Sumeracki, M., & Caviglioli, O. (2018). *Understanding how we learn: A visual guide*. Routledge.

Weissgerber, S. (2019). *Factors Influencing the Effectivity and Usage of Desirable Difficulties in Learning* [University of Kassel]. https://kobra.uni-kassel.de/themes/Mirage2/scripts/mozilla-pdf.js/web/viewer.html?file=/bitstream/handle/123456789/11431/DissertationSophiaChristinWeissgerber.pdf?sequence=8&isAllowed=y#pagemode=thumbs

Welle-Donker, F., van Loenen, B., Keßler, C., Küppers, N., Panek, M., Mansourian, A., Zhao, P., Vancauwenberghe, G., Tomić, H., & Kević, K. (2022). Showcase of Active Learning and Teaching Practices in Spatial Data Infrastructure (SDI) Education. *AGILE: GIScience Series*, *3*, 18.

Wen, W. N. L., & Naim, R. M. (2023) Vocabulary Learning Strategies (VLS) in Second Language Acquisition (SLA): A Review of Literature. *International Journal of Language, Literacy and Translation. 6*. 223-241. 10.36777/ijollt2023.6.2.087.

Williams, M., Mercer, S., & Ryan, S. (2016). *Exploring psychology in language learning and teaching*. Oxford University Press.

Willingham, D. T. (2021). *Why don't students like school?: A cognitive scientist answers questions about how the mind works and what it means for the classroom*. John Wiley & Sons.

Willingham, D. T. (2023). *Outsmart your brain: why learning is hard and how you can make it easy*. Simon and Schuster.

Willis, J. (2007). Review of research: Brain-based teaching strategies for improving students' memory, learning, and test-taking success. *Childhood Education*, *83*(5), 310–315.

Wilson, M. (2009). Responsive writing assessment. Educational Leadership. *Educational Leadership*, *67*(3), 58–62.

Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2012). How and when do students use flashcards? *Memory*, *20*(6), 568–579.

Wittrock, M. C. (1974a). A generative model of mathematics learning. *Journal for Research in Mathematics Education*, 181–196.

Wittrock, M. C. (1974b). Learning as a generative process. *Educational Psychologist*, *11*(2), 87–95. https://doi.org/doi.org/10.1080/00461527409529129

Wittrock, M. C. (1989). Generative processes of comprehension. *Educational Psychologist*, *24*(4), 345–376. https://doi.org/doi.org/10.1207/s15326985ep2404_2

Wittrock, M. C. (1992). Generative learning processes of the brain. *Educational Psychologist*, *27*(4), 531–541.

Wittrock, M. C., & Alesandrini, K. (1990). Generation of summaries and analogies and analytic and holistic abilities. *American Educational Research Journal*, *27*(3), 489-502.

Wood, K. D. (1986). The effect of interspersing questions in text: Evidence for "slicing the task." *Literacy Research and Instruction*, *25*(4), 295–307. https://doi.org/doi.org/10.1080/19388078609557888

Wubalem, A. Y. (2021). Assessing learning transfer and constraining issues in EAP writing practices. *Asian-Pacific Journal of Second and Foreign Language Education*, *6*(1), 1-22.

Yousefi, M. H., & Biria, R. (2018). The effectiveness of L2 vocabulary instruction: A meta-analysis. *Asian-Pacific Journal of Second and Foreign Language Education*, *3*, 1–19.

Yu, S., Jiang, L., & Zhou, N. (2020). Investigating what feedback practices contribute to students' writing motivation and engagement in Chinese EFL context: A large scale study. *Assessing Writing*, *44*, 100451.

Zormpa, E., Brehm, L. E., Hoedemaker, R. S., & Meyer, A. S. (2019). The production effect and the generation effect improve memory in picture naming. *Memory*, *27*(3), 340–352.

Appendices

Appendix 1a.

Informed Consent

Project title: The Generation Effect in the University EFL Classroom

Researchers: John Duplice, Gareth Carrol, and Paul Thompson

Purpose: The purpose of this study is to find out how people learn effectively. If you have any questions about the study, you can contact John Duplice [redacted] or Gareth Carrol [redacted]

Please read the following information and circle a response as necessary:

- I confirm that the purpose of the study has been explained and that I have understood it.
  YES/NO.
- I have had the opportunity to ask questions and they have been successfully answered.
  YES/NO
- I understand that my participation in this study is voluntary and that I am free to withdraw from the study at any time, without giving a reason and without consequence.
  YES/NO
- I understand that data will be collected in an identifiable form, but that it will be treated as confidential and anonymised in any research outputs.
  YES/NO
- I understand that there are no known risks or hazards associated with participating in this study.
  YES/NO
- . I confirm that I have read and understood the above information and that I agree to participate in this study.
  YES/NO

Participant's signature: _____ Date: _____

Participant's Name (in block capitals): _____

Researcher's signature: _____ Date: _____

Appendix 2a: Pre / Post Assessments for Lesson 1

Target Vocabulary Assessment

The words below are the target vocabulary for this lesson. Give each word a score of 1, 2, 3, 4 for your current understanding of each word. If you know the word, please write a sentence using it or the meaning of the word in the blank space.

1) I do not know this word.

2) I have seen the word before, but I am not sure what it means.

3) I understand this word when I see or hear it in a sentence, but I don`t know how to use it.

4) I know this word and can use it in my speaking and writing.

Vocabulary Word	Score 1,2,3,4	Definition or sample sentence using the vocabulary word to show you can use it.
aroma		
beverage		
cluster		
combine		
condensed		
contemporary		
cultivate		
divine		
humid		
odor		
palate		
paradise		
plantation		
rapid		
rate		
soothing		
subtle		
texture		
toxic		
vary		

Appendix 2b: Pre / Post Assessment for Lesson 2

Target Vocabulary Assessment

The words below are the target vocabulary for this lesson. Give each word a score of 1, 2, 3, 4 for your current understanding of each word. If you know the word, please write a sentence using it or the meaning of the word in the blank space.

1) I do not know this word.

2) I have seen the word before, but I am not sure what it means.

3) I understand this word when I see or hear it in a sentence, but I don`t know how to use it.

4) I know this word and can use it in my speaking and writing.

Vocabulary Word	Score 1,2,3,4	Definition or sample sentence using the vocabulary word to show you can use it.
accurate		
analyze		
asteroid		
controversy		
evolve		
factor		
genetic		
genome		
identical		
intellectual		
majority		
mammal		
multiply		
offspring		
pesticide		
regulate		
reinforce		
stricken		
vast		
vegetarian		

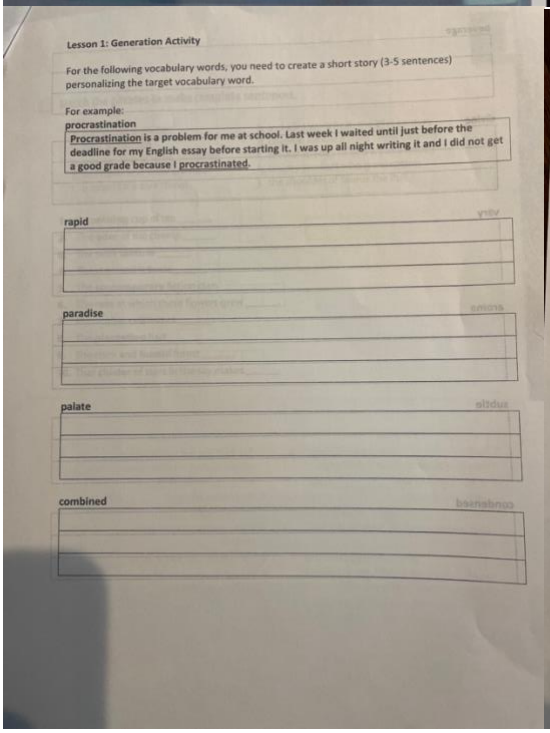
Appendix 3:
Lesson 1 "The History of Chocolate"

UNIT 1

WORD LIST

- **aroma** [əróumə]
  n. An **aroma** is a scent or smell.
  I love the **aroma** of coffee in the morning.
- **beverage** [bévərɪdʒ]
  n. A **beverage** is a drink.
  The waiter brought our **beverages** first. Then he brought our food.
- **cluster** [klʌstər]
  n. A **cluster** of things is a small group of them placed close together.
  She held a large **cluster** of grapes in her hand.
- **combine** [kəmbáɪn]
  v. To **combine** is to join together to make a single thing or group.
  Mina **combined** peanut butter and jelly to make a sandwich.
- **condensed** [kəndénst]
  adj. When a liquid is **condensed**, it is made thicker.
  One way to make a dessert thick and sweet is to use **condensed** milk.
- **contemporary** [kəntémpərèri]
  adj. When something is **contemporary**, it is related to the present time.
  Contemporary scientists have learned quite a bit about DNA.
- **cultivate** [kʌ́ltəvèɪt]
  v. To **cultivate** plants is to care for them and help them grow.
  A research company is **cultivating** new kinds of rice to aid poor countries.
- **divine** [dəváɪn]
  adj. When something is **divine**, it is related to gods.
  Legends say that music was given to humans as a **divine** gift from the gods.
- **humid** [hjúːmɪd]
  adj. When it is **humid**, there is a lot of water in the air.
  It is very **humid** inside a sauna.
- **odor** [óudər]
  n. An **odor** is a very distinct smell.
  He knew there was a leak when he noticed the strong **odor** of natural gas.

- **palate** [pælɪt]
  n. The **palate** is the top part of the mouth.
  You can touch your **palate** with your tongue.
- **paradise** [pærədàɪs]
  n. **Paradise** is the place or condition of happiness where things are perfect.
  My vacation in Hawaii was like being in **paradise**.
- **plantation** [plæntéɪʃən]
  n. A **plantation** is a big farm that only grows certain kinds of crops.
  In the 1800s, there were many cotton **plantations** in the southern US.
- **rapid** [rǽpɪd]
  adj. When something is **rapid**, it moves or changes very quickly.
  His mother was surprised by her son's **rapid** growth.
- **rate** [reɪt]
  n. A **rate** is the speed at which something happens.
  Grass tends to grow at a very slow **rate**.
- **soothing** [sú:ðɪŋ]
  adj. When something is **soothing**, it makes you calm or relaxed.
  The **soothing** music helped the baby fall asleep.
- **subtle** [sʌ́tl]
  adj. When something is **subtle**, it is not easy to see or notice.
  The handsome man has a **subtle** smile.
- **texture** [tékstʃər]
  n. The **texture** of something is the way its surface looks and feels.
  The **texture** of a rock found in the water is typically very smooth.
- **toxic** [tɑ́ksɪk]
  adj. When something is **toxic**, it is poisonous and very dangerous.
  Please check the label to see if the product is **toxic**.
- **vary** [vɛ́əri]
  v. To **vary** means to be different from another thing in size or amount.
  The heights of the people in my class **vary** by a large amount.



Lesson 1: Generation Activity

For the following vocabulary words, you need to create a short story (3-5 sentences) personalizing the target vocabulary word.

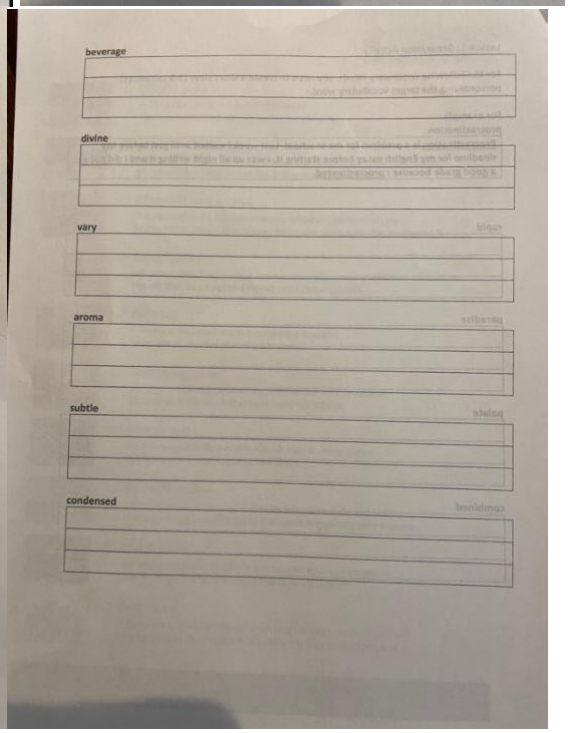
For example:
procrastination
Procrastination is a problem for me at school. Last week I waited until just before the deadline for my English essay before starting it. I was up all night writing it and I did not get a good grade because I procrastinated.

rapid

paradise

palate

combined



beverage

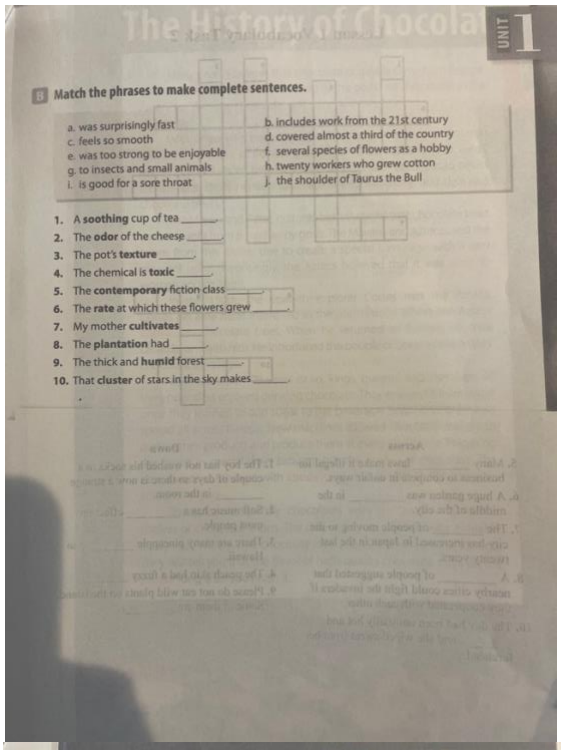
divine

vary

aroma

subtle

condensed

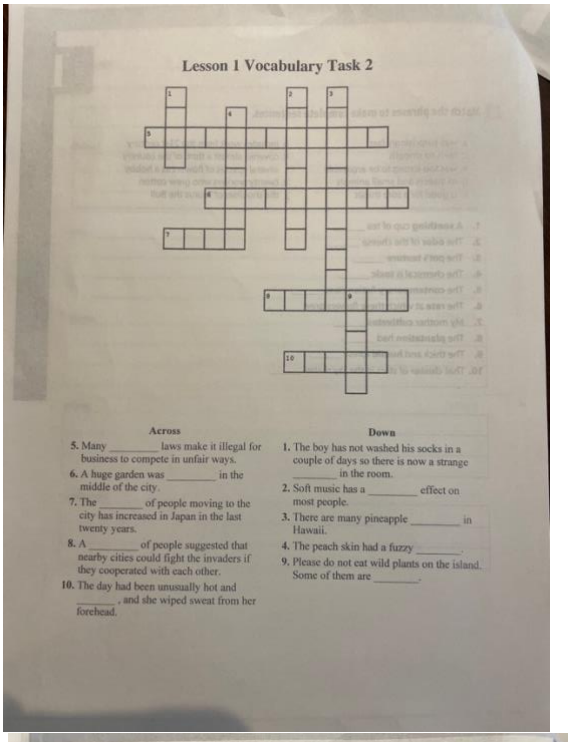


UNIT 1

B **Match the phrases to make complete sentences.**

a. was surprisingly fast	b. includes work from the 21st century
c. feels so smooth	d. covered almost a third of the country
e. was too strong to be enjoyable	f. several species of flowers as a hobby
g. to insects and small animals	h. twenty workers who grew cotton
i. is good for a sore throat	j. the shoulder of Taurus the Bull

1. A **soothing** cup of tea _____.
2. The **odor** of the cheese _____.
3. The pot's **texture** _____.
4. The chemical is **toxic** _____.
5. The **contemporary** fiction class _____.
6. The **rate** at which these flowers grew _____.
7. My mother **cultivates** _____.
8. The **plantation** had _____.
9. The thick and **humid** forest _____.
10. That **cluster** of stars in the sky makes _____.



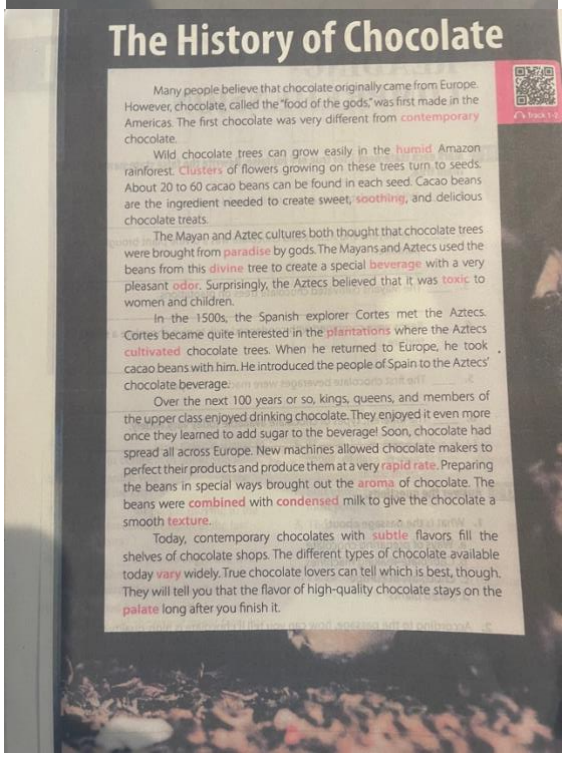
Lesson 1 Vocabulary Task 2

Across

5. Many _____ laws make it illegal for business to compete in unfair ways.
6. A huge garden was _____ in the middle of the city.
7. The _____ of people moving to the city has increased in Japan in the last twenty years.
8. A _____ of people suggested that nearby cities could fight the invaders if they cooperated with each other.
10. The day had been unusually hot and _____, and she wiped sweat from her forehead.

Down

1. The boy has not washed his socks in a couple of days so there is now a strange _____ in the room.
2. Soft music has a _____ effect on most people.
3. There are many pineapple _____ in Hawaii.
4. The peach skin had a fuzzy _____.
9. Please do not eat wild plants on the island. Some of them are _____.



The History of Chocolate

Many people believe that chocolate originally came from Europe. However, chocolate, called the "food of the gods," was first made in the Americas. The first chocolate was very different from contemporary chocolate.

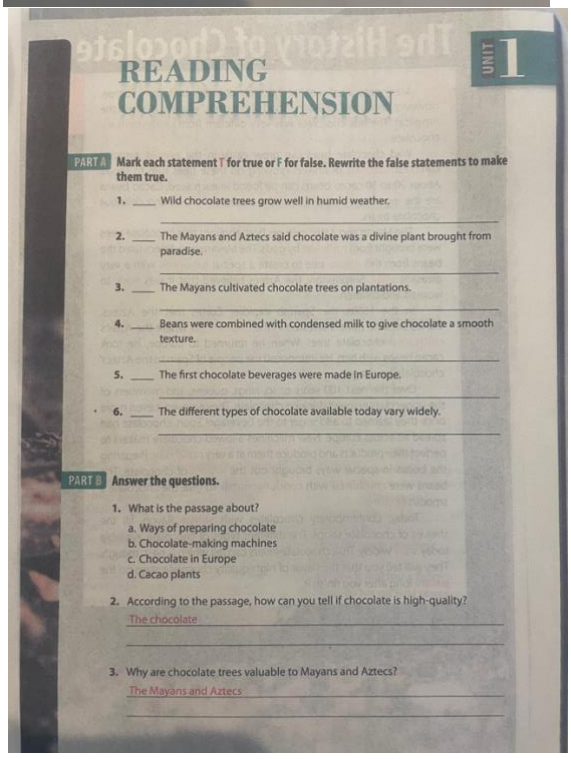
Wild chocolate trees can grow easily in the humid Amazon rainforest. Clusters of flowers growing on these trees turn to seeds. About 20 to 60 cacao beans can be found in each seed. Cacao beans are the ingredient needed to create sweet, soothing, and delicious chocolate treats.

The Mayan and Aztec cultures both thought that chocolate trees were brought from paradise by gods. The Mayans and Aztecs used the beans from this divine tree to create a special beverage with a very pleasant odor. Surprisingly, the Aztecs believed that it was toxic to women and children.

In the 1500s, the Spanish explorer Cortes met the Aztecs. Cortes became quite interested in the plantations where the Aztecs cultivated chocolate trees. When he returned to Europe, he took cacao beans with him. He introduced the people of Spain to the Aztecs' chocolate beverage.

Over the next 100 years or so, kings, queens, and members of the upper class enjoyed drinking chocolate. They enjoyed it even more once they learned to add sugar to the beverage! Soon, chocolate had spread all across Europe. New machines allowed chocolate makers to perfect their products and produce them at a very rapid rate. Preparing the beans in special ways brought out the aroma of chocolate. The beans were combined with condensed milk to give the chocolate a smooth texture.

Today, contemporary chocolates with subtle flavors fill the shelves of chocolate shops. The different types of chocolate available today vary widely. True chocolate lovers can tell which is best, though. They will tell you that the flavor of high-quality chocolate stays on the palate long after you finish it.



UNIT 1

READING COMPREHENSION

PART A **Mark each statement T for true or F for false. Rewrite the false statements to make them true.**

1. ____ Wild chocolate trees grow well in humid weather.
2. ____ The Mayans and Aztecs said chocolate was a divine plant brought from paradise.
3. ____ The Mayans cultivated chocolate trees on plantations.
4. ____ Beans were combined with condensed milk to give chocolate a smooth texture.
5. ____ The first chocolate beverages were made in Europe.
6. ____ The different types of chocolate available today vary widely.

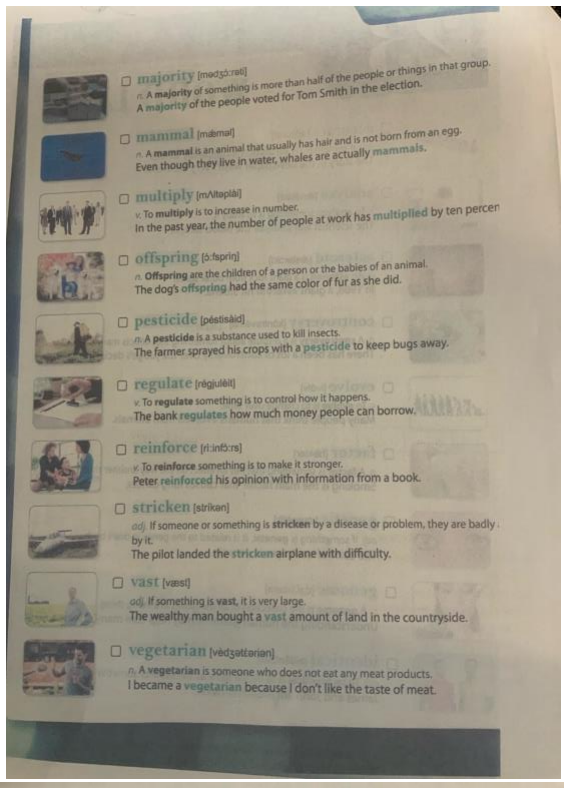
PART B **Answer the questions.**

1. What is the passage about?
   a. Ways of preparing chocolate
   b. Chocolate-making machines
   c. Chocolate in Europe
   d. Cacao plants
2. According to the passage, how can you tell if chocolate is high-quality?
   The chocolate
3. Why are chocolate trees valuable to Mayans and Aztecs?
   The Mayans and Aztecs

Lesson 2: “How the Dinosaurs Really Died”

UNIT 23 WORD LIST

- **accurate** [ækjurət]
  adj. If something is **accurate**, it is completely correct.
  The story in the newspaper wasn't very **accurate**.
- **analyze** [ænəlàiz]
  v. To **analyze** something is to study it.
  The scientist will **analyze** the blood sample.
- **asteroid** [æstərɔ̀id]
  n. An **asteroid** is a giant rock from outer space.
  In 1908, a giant **asteroid** hit Siberia.
- **controversy** [kɑ́ntrəvə̀:rsi]
  n. A **controversy** is a dispute about something that affects many people.
  There has been a lot of **controversy** over the judge's decision.
- **evolve** [iválv]
  v. To **evolve** is to change over time.
  Many people think that humans **evolved** from animals.
- **factor** [fǽktər]
  n. A **factor** is something that has an effect on the way another thing happens.
  Smoking is the main **factor** that causes lung cancer.
- **genetic** [dʒənétik]
  adj. If something is **genetic**, it is related to the genes in one's body.
  The color of one's eyes is **genetic**.
- **genome** [dʒí:noum]
  n. A **genome** is the collection of all the genes in a living thing.
  Understanding the human **genome** may help cure many diseases.
- **identical** [aidéntikəl]
  adj. To be **identical** is to be the same as someone or something else.
  James and John are **identical** twins.
- **intellectual** [intəlékʧuəl]
  n. An **intellectual** is a very smart person.
  We've always considered my Uncle Max to be the **intellectual** of the family.



- **majority** [məʤɔ́:rəti]
  n. A **majority** of something is more than half of the people or things in that group.
  A **majority** of the people voted for Tom Smith in the election.
- **mammal** [mǽməl]
  n. A **mammal** is an animal that usually has hair and is not born from an egg.
  Even though they live in water, whales are actually **mammals**.
- **multiply** [mʌ́ltəplài]
  v. To **multiply** is to increase in number.
  In the past year, the number of people at work has **multiplied** by ten percen
- **offspring** [ɔ́:fspriŋ]
  n. **Offspring** are the children of a person or the babies of an animal.
  The dog's **offspring** had the same color of fur as she did.
- **pesticide** [péstisàid]
  n. A **pesticide** is a substance used to kill insects.
  The farmer sprayed his crops with a **pesticide** to keep bugs away.
- **regulate** [régjulèit]
  v. To **regulate** something is to control how it happens.
  The bank **regulates** how much money people can borrow.
- **reinforce** [rì:infɔ́:rs]
  v. To **reinforce** something is to make it stronger.
  Peter **reinforced** his opinion with information from a book.
- **stricken** [stríkən]
  adj. If someone or something is **stricken** by a disease or problem, they are badly by it.
  The pilot landed the **stricken** airplane with difficulty.
- **vast** [væst]
  adj. If something is **vast**, it is very large.
  The wealthy man bought a **vast** amount of land in the countryside.
- **vegetarian** [vèdʒətɛ́əriən]
  n. A **vegetarian** is someone who does not eat any meat products.
  I became a **vegetarian** because I don't like the taste of meat.

Lesson 2: Generation Activity

For the following vocabulary words, you need to create a short story (3-5 sentences) personalizing the target vocabulary word.

For example:

procrastination

Procrastination is a problem for me at school. Last week I waited until just before the deadline for my English essay before starting it. I was up all night writing it and I did not get a good grade because I procrastinated.

multiply

genetic

accurate

controversy

factor

intellectual

majority

offspring

regulate

vegetarian

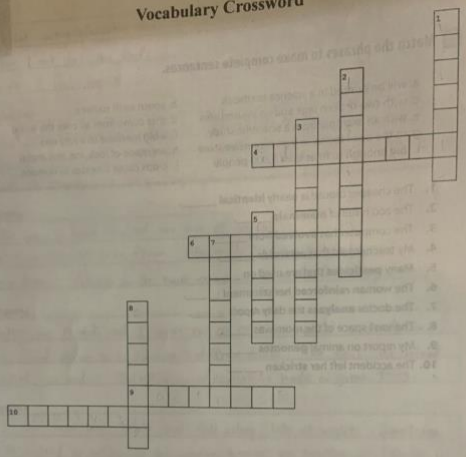
UNIT 23

B **Match the phrases to make complete sentences.**

a. will be printed in a science textbook	b. about each patient
c. with two broken legs and an injured arm	d. that come from all over the world
e. with an example from a scientific study	f. a big machine to a tiny one
g. to the one I found in that expensive store	h. are made of rock, ice, and metal
i. big enough to fit at least 5,000 people	j. crops cause illnesses in humans

1. The cheaper blouse is nearly **identical** _____.
2. The zoo is full of **mammals** _____.
3. The computer has **evolved** from _____.
4. My teacher said that **asteroids** _____.
5. Many **pesticides** that are used on _____.
6. The woman **reinforced** her statement _____.
7. The doctor **analyzes** the daily report _____.
8. The **vast** space of the room was _____.
9. My report on animal **genomes** _____.
10. The accident left her **stricken** _____.

Vocabulary Crossword



Across

4. Twins that look the same are known as _____ Twins.
6. It is believed that an _____ is responsible for the extinction of the dinosaurs.
9. The _____ of space is difficult to comprehend.
10. Scientist _____ the data they collect to learn new things.

Down

1. Monkeys, dogs, and humans are all _____.
2. Farmers us _____ to stop insects from eating plants.
3. The concrete is very strong because it is _____.
5. Many scientists are studying _____ to help cure diseases that were previously incurable.
7. The teacher was left _____ with hearing loss after the accident.
8. When something changes, it _____.

How the Dinosaurs Really Died

Track 23-2

Many scientists and **intellectuals** think that dinosaurs died when an **asteroid** smashed into the Earth millions of years ago. However, recently, there has been some **controversy** over this theory. Some scientists think that it isn't **accurate**. They think that a tiny insect may have been the biggest **factor** in the death of these huge creatures. That insect was the mosquito.

These scientists do think that an asteroid hit the Earth in the time of the dinosaurs. But that wasn't what killed all of them. At that time, insects, including the mosquito, were beginning to **evolve**. Today, we can **regulate** the number of mosquitoes with **pesticides**. But that was impossible millions of years ago. The mosquitoes **multiplied** quickly. And they were certainly not idle. Since there were so many mosquitoes, it was easy for them to bite many of the dinosaurs. When they bit another living thing, the mosquitoes passed along a deadly disease. So the dinosaurs were **stricken** with the disease. The **vast majority** of them—from the **vegetarians** to the meat eaters—died.

To **reinforce** this idea, scientists stress how gradually the dinosaurs died. If an asteroid had killed them, they would have died very quickly. But the number of dinosaurs decreased slowly. In addition, scientists have found **genetic** material of mosquitoes in fossils. This material proves that mosquitoes existed back then. Although there may have been other factors, the dinosaurs died mainly because of disease, the scientists say.

No matter how it happened, the dinosaurs' death had a major impact on other living things. Many dinosaurs ate **mammals**. After the dinosaurs died, mammals were able to evolve and produce **offspring**. Birds also evolved. Scientists have **analyzed** the **genomes** of birds and have discovered that birds have **identical** genetic material to some dinosaurs. So, there may still be dinosaurs among us after all.

144

UNIT 23

READING COMPREHENSION

PART A **Mark each statement T for true or F for false. Rewrite the false statements to make them true.**

1. ___ Some scientists think the asteroid theory isn't accurate.
2. ___ A huge creature may have been the biggest factor in the death of these tiny insects.
3. ___ Today, we can regulate the number of mosquitoes with pesticides.
4. ___ The vast majority of mosquitoes, from the vegetarians to the meat eaters, died.
5. ___ In addition, scientists have found the genetic material of mammals in fossils.
6. ___ Many dinosaurs ate mammals.

PART B **Answer the questions.**

1. What is the passage about?
   a. Running out of food
   b. Several factors that caused dinosaur extinction
   c. Birds descending from dinosaurs
   d. Asteroids coming to Earth
2. What do we do to regulate the number of mosquitoes?
   We use _____
3. How did the mosquitoes spread the deadly disease?
   They bit _____

Appendix 4: Reading Tasks Baseline and Treatments with Embedded Generation Prompts and Comprehension Questions

Baseline Reading

THE MAD HATTER

One morning, Lucas sat outside with his grandfather. They looked past the ***gravel*** road that led to a natural ***reservoir*** on the ***delta***. On the other side of the water, there was a cottage.

"Does a ghost live there?" Lucas asked.

"No, a mad hatter lives there," said his grandfather. Lucas didn't know what a mad hatter was, but the image of a scary man ***haunted*** him.

Later, Lucas went for a walk in the forest. He collected pieces of ***amber*** and ***granite*** that he found on the ground. He looked at the ***moss*** on the trees and watched a bird ***peck*** at the ground. But the forest was like a ***maze***. Soon, Lucas was lost.

Lucas heard somebody behind him. He wanted to run away, but he fell. He had a ***streak*** of blood on his shirt and some ***pebbles*** stuck in his skin. Then a man appeared.

"I will take you home. First, let's get you cleaned up," he said.

Lucas followed him. When they arrived at the cottage, he realized the man was the mad hatter!

He sat down inside. It smelled like ***charcoal***, but it looked like a normal house. The man brought Lucas back some medicine.

"It's a bit old, but it's not ***expired***," the man said.

While Lucas cleaned his cut, the man washed the blood out of his shirt with ***detergent***.

Lucas asked, "Are you a mad hatter?"

The man laughed and replied, "That's a ***euphemism*** for a crazy person. Actually, I'm pretty normal. I'm a ***columnist*** fora newspaper," said the man. He pointed to his ***credentials*** which hung on the wall.

Lucas could hear the ***crickets*** outside. It was getting dark, so he asked, "Could you take me home now?"

The man said yes. Lucas was surprised that people thought the man was crazy. He was actually very ***courteous***. Maybe Lucas should have a more ***liberal*** attitude. Next time, Lucas wouldn't make judgments about people without getting to know them first.

THE MAD HATTER

One morning, Lucas sat outside with his grandfather. They looked past the ***gravel*** road that led to a natural ***reservoir*** on the ***delta***. On the other side of the water, there was a cottage.

"Does a ghost live there?" Lucas asked.

"No, a mad hatter lives there," said his grandfather. Lucas didn't know what a mad hatter was, but the image of a scary man ***haunted*** him.

What do you think a mad hatter is? Use your imagination.	
Why do you think it haunted Lucas?	

Later, Lucas went for a walk in the forest. He collected pieces of ***amber*** and ***granite*** that he found on the ground. He looked at the ***moss*** on the trees and watched a bird ***peck*** at the ground. But the forest was like a ***maze***. Soon, Lucas was lost.

Lucas heard somebody behind him. He wanted to run away, but he fell. He had a ***streak*** of blood on his shirt and some ***pebbles*** stuck in his skin. Then a man appeared.

"I will take you home. First, let's get you cleaned up," he said.

Lucas followed him. When they arrived at the cottage, he realized the man was the mad hatter!

He sat down inside. It smelled like ***charcoal***, but it looked like a normal house. The man brought Lucas back some medicine.

"It's a bit old, but it's not ***expired***," the man said.

While Lucas cleaned his cut, the man washed the blood out of his shirt with ***detergent***.

Lucas asked, "Are you a mad hatter?"

Would you ask if the man was a mad hatter? Why or why not?	

The man laughed and replied, “That’s a ***euphemism*** for a crazy person. Actually, I’m pretty normal. I’m a ***columnist*** for a newspaper,” said the man. He pointed to his ***credentials,*** which hung on the wall.

Lucas could hear the ***crickets*** outside. It was getting dark, so he asked, “Could you take me home now?”

The man said yes. Lucas was surprised that people thought the man was crazy. He was actually very ***courteous***. Maybe Lucas should have a more ***liberal*** attitude. Next time, Lucas wouldn’t make judgments about people without getting to know them first.

Describe how Lucas may have felt after he talked with the man?	
What do you think the moral (lesson learned) of the story is?	

Post-test for both conditions

Comprehension Questions

Reading Comprehension

Part A: Mark each statement T for true or F for false. Rewrite the false statements to make them true.

1. T / F Lucas's thoughts about the house across the reservoir haunted him.

Rewrite false statements to make them true here.	

2. T / F Lucas collected crickets and moss while he walked through the forest.

Rewrite false statements to make them true here.	

3. T / F The bird was pecking at the maze.

Rewrite false statements to make them true here.	

4. T / F Lucas got a streak of blood on his shirt and pebbles in his skin from falling down.

Rewrite false statements to make them true here.	

5. T / F The man's house smelled like food that had expired.

Rewrite false statements to make them true here.	

Part B: Answer the questions.

1. Where was the cottage located in relation to the grandfather's house?

2. What did the man wash Lucas's shirt with?

3. Why did the man point to his credentials on the wall?

4. What did the man say about the term “ mad hatter”?

5. What did the courteous man teach Lucas at the end of the story?

Baseline Reading

THE TENACIOUS INVENTOR

A young student of ***meteorology*** was having a difficult time with an experiment. He was attempting to ***duplicate*** lightning in clouds. He had made a device that could ***simulate*** lightning. It worked by releasing an ***electromagnetic*** pulse into the cloud. This pulse, in turn, ***stimulated*** the ***electrons*** in the cloud's particles. Then the electrons produced lightning.

But his ***meteorological*** experiment had a major ***defect***. He couldn't get the device into the sky.

He had tied it to balloons, but they had burst. He had shot the device from a cannon, but the force of the cannon had damaged it.

"You should give up," his friends told him. "You'll never get that thing into the air."

But his friends' criticisms only ***spurred*** him to try again. The student was very ***innovative***, and at last, he thought that he had an ***innovation*** that would work. He attached wings to the device, and on one ***dreary*** day, when clouds blocked the light of the sun, he started his experiment ***anew***.

He placed the device on a rocket and ***launched*** it into the sky. The ***propulsion*** of the rocket carried the device high into the air. The rocket ***accelerated*** into the clouds and then released the device. It ***glided*** on its wings through the clouds, and when it ***penetrated*** the center of a large black cloud, it emitted the electromagnetic pulse. And just as he had predicted, lightning shot from the cloud!

He called his professors, and the next day they came to watch. He successfully duplicated the experiment. His teachers were extremely impressed and called the student and his invention ***ingenious***.

The student was given many awards and became a famous inventor. He had not given up. He had remained ***tenacious*** and succeeded.

THE TENACIOUS INVENTOR

A young student of ***meteorology*** was having a difficult time with an experiment. He was attempting to ***duplicate*** lightning in clouds. He had made a device that could ***simulate*** lightning. It worked by releasing an ***electromagnetic*** pulse into the cloud. This pulse, in turn, ***stimulated*** the ***electrons*** in the cloud's particles. Then the electrons produced lightning.

But his ***meteorological*** experiment had a major ***defect***. He couldn't get the device into the sky.

Describe what the device looks like in your mind. Use your imagination.	
Why do you think this student is doing the experiment?	

He had tied it to balloons, but they had burst. He had shot the device from a cannon, but the force of the cannon had damaged it.

"You should give up," his friends told him. "You'll never get that thing into the air."

But his friends' criticisms only ***spurred*** him to try again. The student was very ***innovative***, and at last, he thought that he had an ***innovation*** that would work. He attached wings to the device, and on one ***dreary*** day, when clouds blocked the light of the sun, he started his experiment ***anew***.

Why do you think he didn`t give up?	

He placed the device on a rocket and ***launched*** it into the sky. The ***propulsion*** of the rocket carried the device high into the air. The rocket ***accelerated*** into the clouds and then released the device. It ***glided*** on its wings through the clouds, and when it ***penetrated*** the center of a large black cloud, it emitted the electromagnetic pulse. And just as he had predicted, lightning shot from the cloud!

He called his professors, and the next day they came to watch. He successfully duplicated the experiment. His teachers were extremely impressed and called the student and his invention ***ingenious***.

The student was given many awards and became a famous inventor. He had not given up. He had remained ***tenacious*** and succeeded.

Describe (in your opinion) how the student while his	

teachers were watching experiment.	
Describe the moral (lesson learned) of the story.	

Post-test for both conditions

Reading Comprehension

Part A: Mark each statement T for true or F for false. Rewrite the false statements to make them true.

1. T / F The student of meteorology had bought a device that simulated lightning in clouds.

Rewrite false statements to make them true here.	

2. T / F The electromagnetic pulse stimulated the electrons in the cloud's particles.

Rewrite false statements to make them true here.	

3. T / F The student's friends' criticisms spurred him to try his experiment anew.

Rewrite false statements to make them true here.	

4. T / F It was a dreary day when the device glided into the clouds.

Rewrite false statements to make them true here.	

5. T / F The propulsion of the rocket accelerated the speed of the lightning.

Rewrite false statements to make them true here.	

Part B: Answer the questions.

1. What was the defect of the student's meteorological experiment?

2. What innovation did the innovative student use to launch his device into the clouds?

3. For whom did the student duplicate his ingenious experiment?

4. What did the device do when it finally penetrated the center of a large black cloud?

5. What happened to the student because he was tenacious?

Baseline Reading

THE FOSSIL HUNTERS

Tim and Dean were great fossil hunters. They were the very best at finding dinosaur bones. Although Tim and Dean were quite similar, they were ***outright*** enemies. The two men got into ***vicious*** arguments all the time. They couldn't ***coexist*** peacefully because their ***egos*** were too large. Tim thought he was the best fossil hunter, while Dean was sure that he was much better than Tim.

One day, Tim was searching for fossils on the ***periphery*** of the city when he discovered a huge bone. He had never seen anything like it! He took his ***shovel*** and carefully ***excavated*** the dirt around it. As he dug, he uncovered more ***jagged*** bones. He realized that he had found an entire dinosaur ***skeleton***! Tim couldn't ***conceive*** a plan to remove the huge skeleton all by himself. Such an ***endeavor*** would be too ***arduous***. He needed help. He tried to think of people who would be capable of helping him remove the skeleton without breaking it. The only person Tim could think of was Dean, his enemy.

Tim ran into the city to find Dean. Tim found him and said, " Dean, I've found the ***skeletal*** remains of a huge ***terrestrial*** animal. But I can't get the skeleton out by myself. Will you please help me?

Dean thought that Tim's claim might be ***dubious***. He replied, " If you're serious about the skeleton, I'll help."

Tim excitedly showed Dean the skeleton's ***locale***. They worked together to carefully remove each bone. And to keep the bones together, they tied them with ***elastic*** strips. When they were finished, they had ***attained*** a perfect skeleton. They used ***plaster*** to make a ***mold*** of the dinosaur's skull. They ***engraved*** their initials into it and gave it to the curator of a local museum.

Tim and Dean found out that they could work very well together. They decided to end their feud and become friends. By combining their talents, the men became even greater than they were before.

THE FOSSIL HUNTERS

Tim and Dean were great fossil hunters. They were the very best at finding dinosaur bones. Although Tim and Dean were quite similar, they were ***outright*** enemies. The two men got into ***vicious*** arguments all the time. They couldn't ***coexist*** peacefully because their ***egos*** were too large. Tim thought he was the best fossil hunter, while Dean was sure that he was much better than Tim.

In your own words, how would you describe Tim and Dean?	
What is the problem?	

One day, Tim was searching for fossils on the ***periphery*** of the city when he discovered a huge bone. He had never seen anything like it! He took his ***shovel*** and carefully ***excavated*** the dirt around it. As he dug, he uncovered more ***jagged*** bones. He realized that he had found an entire dinosaur ***skeleton***! Tim couldn't ***conceive*** a plan to remove the huge skeleton all by himself. Such an ***endeavor*** would be too ***arduous***. He needed help. He tried to think of people who would be capable of helping him remove the skeleton without breaking it. The only person Tim could think of was Dean, his enemy.

Describe how you would feel if you were in Tim`s situation and you had to ask for Dean`s help?	

Tim ran into the city to find Dean. Tim found him and said, " Dean, I've found the ***skeletal*** remains of a huge ***terrestrial*** animal. But I can't get the skeleton out by myself. Will you please help me?

Dean thought that Tim's claim might be ***dubious***. He replied, " If you're serious about the skeleton, I'll help."

Tim excitedly showed Dean the skeleton's ***locale***. They worked together to carefully remove each bone. And to keep the bones together, they tied them with ***elastic*** strips. When they were finished, they had ***attained*** a perfect skeleton. They used ***plaster*** to make a ***mold*** of the dinosaur's skull. They ***engraved*** their initials into it and gave it to the curator of a local museum.

Tim and Dean found out that they could work very well together. They decided to end their feud and become friends. By combining their talents, the men became even greater than they were before.

Describes (in your opinion) how Dean felt when Tim asked him for help.	

Describe the main moral (lesson learned) of the story.	

Post-test for both conditions

Part A: Mark each statement T for true or F for false. Rewrite the false statements to make them true.

1. T / F Tim and Dean were outright enemies who got into vicious arguments.

Rewrite false statements to make them true here.	

2. T / F Tim and Dean removed the jagged bones and used elastic to attain them.

Rewrite false statements to make them true here.	

3. T / F Tim couldn't conceive a plan to remove the bones because the endeavor would be too arduous.

Rewrite false statements to make them true here.	

4. T / F Dean thought that Tim's ego might be dubious.

Rewrite false statements to make them true here.	

5. T / F When Tim excavated the land, he uncovered many engraved bones.

Rewrite false statements to make them true here.	

Part B: Answer the questions.

1. Why couldn't Tim and Dean coexist peacefully?
2. What did Tim use his shovel to do?
3. What was Tim doing on the periphery of the city?
4. What did the fossil hunters do to the plaster mold before they gave it to the curator?
5. What did Dean say before he went to the locale of the terrestrial animal's skeletal remains?

Appendices (for T/F scores) (4a-4c)

Appendix 4a: Covariate Summary for Gender, True / False Scores

Names	Effect	Estimate	SE	exp(B)	95% Exp(B) Confidence Interval		z	p
					Lower	Upper		
(Intercept)	(Intercept)	0.6111	0.0960	1.843	1.526	2.22	6.364	< .001
Treatment1	Gen_L1 - No_Gen	-0.2090	0.2530	0.811	0.494	1.33	-0.826	0.409
Treatment2	Gen_L2 - No_Gen	-0.4036	0.3056	0.668	0.367	1.22	-1.321	0.187
Gender1	Male - Female	-0.0478	0.1876	0.953	0.660	1.38	-0.255	0.799
Treatment1 ✻ Gender1	Gen_L1 - No_Gen ✻ Male - Female	0.6553	0.4935	1.926	0.732	5.07	1.328	0.184
Treatment2 ✻ Gender1	Gen_L2 - No_Gen ✻ Male - Female	-0.3147	0.5989	0.730	0.226	2.36	-0.525	0.599

Appendix 4b: Covariate Summary for Group, True / False Scores

Names	Effect	Estimate	SE	exp(B)	95% Exp(B) Confidence Interval		z	p	
					Lower	Upper			
(Intercept)	(Intercept)	0.6320	0.101	1.8813	1.5426	2.294	6.240	< .001	
Treatment1	Gen_L1 - No_Gen	-0.3375	0.257	0.7136	0.4312	1.181	-1.313	0.189	
Treatment2	Gen_L2 - No_Gen	-0.5590	0.263	0.5718	0.3417	0.957	-2.128	0.033	*
Group1	B - A	-0.2029	0.220	0.8164	0.5304	1.257	-0.922	0.357	
Group2	C - A	-0.0850	0.258	0.9185	0.5539	1.523	-0.329	0.742	
Treatment1 ✻ Group1	Gen_L1 - No_Gen ✻ B - A	0.2111	0.535	1.2350	0.4327	3.525	0.395	0.693	
Treatment2 ✻ Group1	Gen_L2 - No_Gen ✻ B - A	-2.1526	0.557	0.1162	0.0390	0.346	-3.864	< .001	*
Treatment1 ✻ Group2	Gen_L1 - No_Gen ✻ C - A	-1.5408	0.648	0.2142	0.0602	0.762	-2.379	0.017	*
Treatment2 ✻ Group2	Gen_L2 - No_Gen ✻ C - A	-3.0612	0.678	0.0468	0.0124	0.177	-4.513	< .001	*

Appendix 4c: Covariate Summary for Lesson (Study 2), True / False Scores

					95% Exp(B) Confidence Interval			
Names	**Effect**	**Estimate**	**SE**	**exp(B)**	**Lower**	**Upper**	**z**	**p**
(Intercept)	(Intercept)	0.632	0.101	1.881	1.543	2.294	6.240	< .001
Treatment1	Gen_L1 - No_Gen	-0.337	0.257	0.714	0.431	1.181	-1.313	0.189
Treatment2	Gen_L2 - No_Gen	-0.559	0.263	0.572	0.342	0.957	-2.128	0.033
Lesson1	Mad Hatter - Fossil Hunters	0.211	0.224	1.235	0.795	1.916	0.939	0.348
Lesson2	Tenacious Inventor - Fossil Hunters	1.301	0.256	3.675	2.224	6.071	5.082	< .001 *
Treatment1 ✻ Lesson1	Gen_L1 - No_Gen ✻ Mad Hatter - Fossil Hunters	-0.353	0.551	0.703	0.239	2.070	-0.640	0.522
Treatment2 ✻ Lesson1	Gen_L2 - No_Gen ✻ Mad Hatter - Fossil Hunters	-0.348	0.546	0.706	0.242	2.058	-0.638	0.523
Treatment1 ✻ Lesson2	Gen_L1 - No_Gen ✻ Tenacious Inventor - Fossil Hunters	-0.702	0.661	0.496	0.136	1.809	-1.062	0.288
Treatment2 ✻ Lesson2	Gen_L2 - No_Gen ✻ Tenacious Inventor - Fossil Hunters	0.259	0.648	1.296	0.364	4.616	0.400	0.689

Appendices (for Explanatory Scores) (5a-5c)

Appendix 5a: Covariate Summary for Gender (Study 2), Explanatory Scores

					95% Exp(B) Confidence Interval			
Names	**Effect**	**Estimate**	**SE**	**exp(B)**	**Lower**	**Upper**	**z**	**p**
(Intercept)	(Intercept)	0.6111	0.0960	1.843	1.526	2.22	6.364	< .001
Treatment1	Gen_L1 - No_Gen	-0.2090	0.2530	0.811	0.494	1.33	-0.826	0.409
Treatment2	Gen_L2 - No_Gen	-0.4036	0.3056	0.668	0.367	1.22	-1.321	0.187
Gender1	Male - Female	-0.0478	0.1876	0.953	0.660	1.38	-0.255	0.799
Treatment1 ✻ Gender1	Gen_L1 - No_Gen ✻ Male - Female	0.6553	0.4935	1.926	0.732	5.07	1.328	0.184
Treatment2 ✻ Gender1	Gen_L2 - No_Gen ✻ Male - Female	-0.3147	0.5989	0.730	0.226	2.36	-0.525	0.599

Appendix 5b*: Covariate Summary for Group (Study 2) Explanatory Scores*

					95% Exp(B) Confidence Interval			
Names	**Effect**	**Estimate**	**SE**	**exp(B)**	**Lower**	**Upper**	**z**	**p**
(Intercept)	(Intercept)	0.6320	0.101	1.8813	1.5426	2.294	6.240	< .001
Treatment1	Gen_L1 - No_Gen	-0.3375	0.257	0.7136	0.4312	1.181	-1.313	0.189
Treatment2	Gen_L2 - No_Gen	-0.5590	0.263	0.5718	0.3417	0.957	-2.128	0.033
Goup1	B - A	-0.2029	0.220	0.8164	0.5304	1.257	-0.922	0.357
Group2	C - A	-0.0850	0.258	0.9185	0.5539	1.523	-0.329	0.742
Treatment1 ✻ Group1	Gen_L1 - No_Gen ✻ B - A	0.2111	0.535	1.2350	0.4327	3.525	0.395	0.693
Treatment2 ✻ Group1	Gen_L2 - No_Gen ✻ B - A	-2.1526	0.557	0.1162	0.0390	0.346	-3.864	< .001
Treatment1 ✻ Group2	Gen_L1 - No_Gen ✻ C - A	-1.5408	0.648	0.2142	0.0602	0.762	-2.379	0.017
Treatment2 ✻ Group2	Gen_L2 - No_Gen ✻ C - A	-3.0612	0.678	0.0468	0.0124	0.177	-4.513	< .001

Treatment2 ✻ Group1	Gen_L2 - No_Gen ✻ B - A	-0.4857	0.1838	-0.846	-0.12552	83.5	-2.643	0.010 *
Treatment1 ✻ Group2	Gen_L1 - No_Gen ✻ C - A	0.0238	0.2256	-0.418	0.46590	45.2	0.106	0.916
Treatment2 ✻ Group2	Gen_L2 - No_Gen ✻ C - A	-0.2143	0.2077	-0.621	0.19288	83.5	-1.032	0.305

Appendix 5c: Covariate Summary of Lesson for Explanatory Scores

Names	Effect	Estimate	SE	95% Confidence Interval		df	t	p	
				Lower	Upper				
(Intercept)	(Intercept)	1.4667	0.0491	1.3704	1.5630	34.4	29.849	< .001	
Treatment1	Gen_L1 - No_Gen	-0.0683	0.0887	-0.2421	0.1056	45.2	-0.770	0.446	
Treatment2	Gen_L2 - No_Gen	-0.2190	0.0817	-0.3791	-0.0590	83.5	-2.682	0.009	*
Lesson1	Mad Hatter - Fossil Hunters	0.0825	0.0893	-0.0925	0.2575	73.6	0.924	0.358	
Lesson2	Tenacious Inventor - Fossil Hunters	0.0349	0.0884	-0.1383	0.2082	83.1	0.395	0.694	
Treatment1 ✻ Lesson1	Gen_L1 - No_Gen ✻ Mad Hatter - Fossil Hunters	0.1746	0.2418	-0.2992	0.6485	52.8	0.722	0.473	
Treatment2 ✻ Lesson1	Gen_L2 - No_Gen ✻ Mad Hatter - Fossil Hunters	0.1159	0.2479	-0.3699	0.6017	47.4	0.467	0.642	
Treatment1 ✻ Lesson2	Gen_L1 - No_Gen ✻ Tenacious Inventor - Fossil Hunters	-0.0270	0.2554	-0.5275	0.4735	50.1	-0.106	0.916	
Treatment2 ✻ Lesson2	Gen_L2 - No_Gen ✻ Tenacious Inventor - Fossil Hunters	0.6841	0.2410	0.2117	1.1566	51.7	2.838	0.006	*

Appendix 6a: Writing Checklist with Generation Tasks

Post Writing Checklist

When you finish writing, go through this checklist to make sure you completed or correctly did each of the points. You MUST mark each of the points with "ok" in the space ( ) provided. Make sure you do this one point at a time, and not just mark "OK" without checking. After you go through and check each of these points, write a few sentences discussing how you made sure it was correctly done or how you fixed it needed to be changed. DO NOT just write, "It was correct / no mistakes."

Example:

(*OK*) The margins are the correct width.

I checked the margins` width, and they were correctly done. I am confident it is correct because the margins are the same width as instructed in Becoming a Better Writer.

This completed checklist must be included with your writing assignment.

( ) My essay has a specific thesis statement and topic sentences for supporting paragraphs.

( ) I followed the correct APA formatting guidelines (e.g., spacing, margin width, font size, etc.).

( ) I understand the purpose of my paper and I successfully accomplished it.

( ) I thoroughly checked for spelling, punctuation, capitalization, and simple grammar mistakes.

( ) I read my paper outload to myself to make sure the writing is clear.

( ) Borrowed content was properly cited in the essay.

( ) The purpose of my paper matches the assignment.

( ) My reference page includes a proper APA reference for every in-text citation in my essay.

( ) I thoroughly read through comments made by peers during the peer-review process and reflected on them when revising my essay.

( ) I checked to make sure I used the correct tense in my writing.

Write a short paragraph explaining what points you did correctly and incorrectly in your writing. Specify which of the above points gave you trouble and how you went about fixing them. At the bottom of the paragraph, type how long it took you to complete the checklist and write the checklist paragraph (for example, 20 minutes).

Type the amount of time it took to do the checklist here:

Appendix 6b: Writing Checklist without Generation Tasks

Post Writing Checklist

When you finish writing, go through this checklist to make sure you completed or correctly did each of the points. You MUST mark each of the points with "ok" in the space ( ) provided. Make sure you do this one point at a time and do not just mark "OK" without really checking. After you go through and check each of these points, write a few sentences discussing how you made sure it was correctly done or how you fixed it needed to be changed. DO NOT just write, "It was correct / no mistakes."

Example:

(*OK*) The margins are the correct width.

I checked the margins` width, and they were correctly done. I am confident it is correct because the margins are the same width as instructed in Becoming a Better Writer.

This completed checklist must be included with your writing assignment.

( ) My essay has a specific thesis statement and topic sentences for supporting paragraphs.

( ) I followed the correct APA formatting guidelines (e.g., spacing, margin width, font size, etc.).

( ) I understand the purpose of my paper and I successfully accomplished it.

( ) I thoroughly checked for spelling, punctuation, capitalization, and simple grammar mistakes.

( ) I read my paper outload to myself to make sure the writing is clear.

( ) Borrowed content was properly cited in the essay.

( ) The purpose of my paper matches the assignment.

( ) My reference page includes a proper APA reference for every in-text citation in my essay.

( ) I thoroughly read through comments made by peers during the peer-review process and reflected on them when revising my essay.

( ) I checked to make sure I used the correct tense in my writing.

Type the amount of time it took to do the checklist here:

Appendix 6c: Post Hoc Comparisons of Formatting - Treatment ✻ Essay with Pairwise Comparisons of Essays

Comparison									
Treatment	**Essay**		**Treatment**	**Essay**	**Difference**	**SE**	**t**	**df**	**$p_{bonferroni}$**
Gen	1	-	Gen	2	-0.3529	0.194	-1.817	87.0	1.000
Gen	1	-	Gen	3	-0.0588	0.194	-0.303	87.0	1.000
Gen	1	-	Gen	4	-0.1176	0.194	-0.606	87.0	1.000
Gen	1	-	NoGen	2	0.3067	0.235	1.307	85.9	1.000
Gen	1	-	NoGen	3	0.3067	0.235	1.307	85.9	1.000
Gen	1	-	NoGen	4	0.0924	0.235	0.394	85.9	1.000
Gen	2	-	Gen	3	0.2941	0.194	1.515	87.0	1.000
Gen	2	-	Gen	4	0.2353	0.194	1.212	87.0	1.000
Gen	2	-	NoGen	3	0.6597	0.235	2.810	85.9	0.172
Gen	2	-	NoGen	4	0.4454	0.235	1.897	85.9	1.000
Gen	3	-	Gen	4	-0.0588	0.194	-0.303	87.0	1.000
Gen	3	-	NoGen	4	0.1513	0.235	0.644	85.9	1.000
NoGen	1	-	Gen	1	-0.3782	0.235	-1.611	85.9	1.000
NoGen	1	-	Gen	2	-0.7311	0.235	-3.114	85.9	0.070
NoGen	1	-	Gen	3	-0.4370	0.235	-1.861	85.9	1.000
NoGen	1	-	Gen	4	-0.4958	0.235	-2.112	85.9	1.000
NoGen	1	-	NoGen	2	-0.0714	0.214	-0.334	87.0	1.000
NoGen	1	-	NoGen	3	-0.0714	0.214	-0.334	87.0	1.000
NoGen	1	-	NoGen	4	-0.2857	0.214	-1.335	87.0	1.000
NoGen	2	-	Gen	2	-0.6597	0.235	-2.810	85.9	0.172
NoGen	2	-	Gen	3	-0.3655	0.235	-1.557	85.9	1.000
NoGen	2	-	Gen	4	-0.4244	0.235	-1.808	85.9	1.000
NoGen	2	-	NoGen	3	2.52e-17	0.214	1.18e-16	87.0	1.000
NoGen	2	-	NoGen	4	-0.2143	0.214	-1.001	87.0	1.000
NoGen	3	-	Gen	3	-0.3655	0.235	-1.557	85.9	1.000
NoGen	3	-	Gen	4	-0.4244	0.235	-1.808	85.9	1.000
NoGen	3	-	NoGen	4	-0.2143	0.214	-1.001	87.0	1.000
NoGen	4	-	Gen	4	-0.2101	0.235	-0.895	85.9	1.000

Appendix 6d: *Post Hoc Comparisons of GSPC - Treatment ✲ Essay with Pairwise Comparisons of Essays*

Comparison									
Treatment	**Essay**		**Treatment**	**Essay**	**Difference**	**SE**	**t**	**df**	$p_{bonferroni}$
Gen	1	-	Gen	2	0.1765	0.128	1.376	87.0	1.000
Gen	1	-	Gen	3	0.0588	0.128	0.459	87.0	1.000
Gen	1	-	Gen	4	-0.1176	0.128	-0.917	87.0	1.000
Gen	1	-	NoGen	2	0.5420	0.176	3.071	71.4	0.084
Gen	1	-	NoGen	3	0.6134	0.176	3.476	71.4	0.024
Gen	1	-	NoGen	4	0.5420	0.176	3.071	71.4	0.084
Gen	2	-	Gen	3	-0.1176	0.128	-0.917	87.0	1.000
Gen	2	-	Gen	4	-0.2941	0.128	-2.293	87.0	0.679
Gen	2	-	NoGen	3	0.4370	0.176	2.476	71.4	0.439
Gen	2	-	NoGen	4	0.3655	0.176	2.071	71.4	1.000
Gen	3	-	Gen	4	-0.1765	0.128	-1.376	87.0	1.000
Gen	3	-	NoGen	4	0.4832	0.176	2.738	71.4	0.218
NoGen	1	-	Gen	1	-0.6134	0.176	-3.476	71.4	0.024
NoGen	1	-	Gen	2	-0.4370	0.176	-2.476	71.4	0.439
NoGen	1	-	Gen	3	-0.5546	0.176	-3.143	71.4	0.068
NoGen	1	-	Gen	4	-0.7311	0.176	-4.142	71.4	0.003
NoGen	1	-	NoGen	2	-0.0714	0.141	-0.505	87.0	1.000
NoGen	1	-	NoGen	3	-5.55e−17	0.141	-3.93e−16	87.0	1.000
NoGen	1	-	NoGen	4	-0.0714	0.141	-0.505	87.0	1.000
NoGen	2	-	Gen	2	-0.3655	0.176	-2.071	71.4	1.000
NoGen	2	-	Gen	3	-0.4832	0.176	-2.738	71.4	0.218
NoGen	2	-	Gen	4	-0.6597	0.176	-3.738	71.4	0.010
NoGen	2	-	NoGen	3	0.0714	0.141	0.505	87.0	1.000
NoGen	2	-	NoGen	4	1.73e-17	0.141	1.23e-16	87.0	1.000
NoGen	3	-	Gen	3	-0.5546	0.176	-3.143	71.4	0.068
NoGen	3	-	Gen	4	-0.7311	0.176	-4.142	71.4	0.003
NoGen	3	-	NoGen	4	-0.0714	0.141	-0.505	87.0	1.000
NoGen	4	-	Gen	4	-0.6597	0.176	-3.738	71.4	0.010

Appendix 6e: ***Post Hoc Comparisons of Content - Treatment ✻ Essay with Pairwise Comparisons of Essays***

Post Hoc Comparisons - Treatment ✻ Essay

Comparison									
Treatment	**Essay**		**Treatment**	**Essay**	**Difference**	**SE**	**t**	**df**	**$p_{bonferroni}$**
Gen	1	-	Gen	2	-0.0588	0.132	-0.445	87.0	1.000
Gen	1	-	Gen	3	-0.1176	0.132	-0.890	87.0	1.000
Gen	1	-	Gen	4	-0.1765	0.132	-1.335	87.0	1.000
Gen	1	-	NoGen	2	0.4706	0.170	2.761	84.5	0.198
Gen	1	-	NoGen	3	0.2563	0.170	1.504	84.5	1.000
Gen	1	-	NoGen	4	0.2563	0.170	1.504	84.5	1.000
Gen	2	-	Gen	3	-0.0588	0.132	-0.445	87.0	1.000
Gen	2	-	Gen	4	-0.1176	0.132	-0.890	87.0	1.000
Gen	2	-	NoGen	3	0.3151	0.170	1.849	84.5	1.000
Gen	2	-	NoGen	4	0.3151	0.170	1.849	84.5	1.000
Gen	3	-	Gen	4	-0.0588	0.132	-0.445	87.0	1.000
Gen	3	-	NoGen	4	0.3739	0.170	2.194	84.5	0.867
NoGen	1	-	Gen	1	-0.6134	0.170	-3.600	84.5	0.015
NoGen	1	-	Gen	2	-0.6723	0.170	-3.945	84.5	0.005
NoGen	1	-	Gen	3	-0.7311	0.170	-4.290	84.5	0.001
NoGen	1	-	Gen	4	-0.7899	0.170	-4.635	84.5	< .001
NoGen	1	-	NoGen	2	-0.1429	0.146	-0.981	87.0	1.000
NoGen	1	-	NoGen	3	-0.3571	0.146	-2.452	87.0	0.454
NoGen	1	-	NoGen	4	-0.3571	0.146	-2.452	87.0	0.454
NoGen	2	-	Gen	2	-0.5294	0.170	-3.106	84.5	0.072
NoGen	2	-	Gen	3	-0.5882	0.170	-3.452	84.5	0.024
NoGen	2	-	Gen	4	-0.6471	0.170	-3.797	84.5	0.008
NoGen	2	-	NoGen	3	-0.2143	0.146	-1.471	87.0	1.000
NoGen	2	-	NoGen	4	-0.2143	0.146	-1.471	87.0	1.000
NoGen	3	-	Gen	3	-0.3739	0.170	-2.194	84.5	0.867
NoGen	3	-	Gen	4	-0.4328	0.170	-2.539	84.5	0.362
NoGen	3	-	NoGen	4	-4.16e−17	0.146	-2.86e−16	87.0	1.000
NoGen	4	-	Gen	4	-0.4328	0.170	-2.539	84.5	0.362